2020

Genomic and Culturomic Analysis of Gut Microbiota Function and Salmonella Enterica Expansion

Gavin Fenske
South Dakota State University

GENOMIC AND CULTUROMIC ANALYSIS OF GUT MICROBIOTA FUNCTION AND SALMONELLA ENTERICA EXPANSION

BY

GAVIN FENSKE

A dissertation submitted in partial fulfillment of the requirements for the

Doctor of Philosophy

Major in Biological Sciences

Specialization in Veterinary Microbiology

South Dakota State University

2020

DISSERTATION ACCEPTANCE PAGE

Gavin Fenske

This dissertation is approved as a creditable and independent investigation by a candidate for the Doctor of Philosophy degree and is acceptable for meeting the dissertation requirements for this degree. Acceptance of this does not imply that the conclusions reached by the candidate are necessarily the conclusions of the major department.

JOY SCARIA
Advisor Date

Jane Hennings
Department Head Date

Dean, Graduate School Date

ACKNOWLEDGEMENTS

Non Impediti Ratione Cogitationis

First, I would like to acknowledge Kassidy Weathers. Without her support none of this work would have been possible. Secondly, I would like to thank Dr. Joy Scaria. When I first came to his lab, I had an interest in science but did not know how much hard work is required to practice it. I know consider myself a scientist, and without Joy's mentorship, that would not be possible.

CONTENTS

ABSTRACT

GENOMIC AND CULTUROMIC ANALYSIS OF GUT MICROBIOTA FUNCTION AND SALMONELLA ENTERICA EXPANSION

GAVIN FENSKE

2020

Enteric bacteria that are resident in the hindgut of mammals are critical in immune development, digestion, and colonization resistance against pathogens. One of the major pathogens that gut commensals provide resistance against is Salmonella enterica, a major foodborne pathogen capable of infecting almost every warm-blooded animal. Given the interplay between pathogens and commensals in the gut lumen, the gut microbiota of pigs was studied by combining two disparate techniques: shotgun metagenomics and high throughput culturomics. Metagenomics readily identifies major taxa present in samples and can give an estimation to total genetic catalogue from an environment. However, many rare or low abundance taxa were retrieved in culture that were not reliably obtained from metagenomics. Major gene pathways recovered from culture isolates were absent from metagenomics. In addition to studying the gut microbiota, two genomics studies were conducted on S. enterica. The first study was to establish and investigate the genomic population structure of a bovine-adapted serovar, *S. enterica* Dublin. The serovar is a primary pathogen of cattle and can establish carrier states with the pathogen being shed intermittently in the feces and milk. It was observed that the core and ancillary

genomes are strongly influence by region of isolation. Virulence and prophage content were homogenous among the geographic clades. However, US genomes yielded multiple antibiotic resistance genes mediated by an IncA/C2 plasmid. Antibiotic resistance was not common in isolates outside of the US. The second S. enterica genome investigation was to identify co-occurrence between metal and antibiotic resistance. Co-occurrence between the genotypes was identified, but is isolated to one clade of *S. enterica* I 4,[5],12:i:-. However, multiple serovars contain resistance to copper and silver, which may permit the expansion into novel niches as metal use continues to rise in medicine and agriculture.

CHAPTER 1: Integration of Culture-Dependent and Independent Methods Provides a More Coherent Picture of the Pig Gut Microbiome

INTRODUCTION

The microbiome in the hindgut of mammals has been associated with feed conversion efficiency [1], pathogen exclusion [2], and the production of metabolites that directly influence host signaling pathways [3]. It has become clear in recent years, that the microbiome has a drastic impact on host health. Many current methods to study the swine microbiome are based upon dietary intervention [4, 5]. That is, a dietary substrate is introduced to the animal and an effect on microbiome composition, typically 16S rRNA analysis, is measured. As sequencing costs have decreased, more studies have relied upon culture-independent methods to explore the microbiota of pigs. However, culture-independent experiments cannot provide mechanistic information about the factors influencing bacterial communities. We attempt to bridge the schism between culture-dependent and independent methods by incorporating both to investigate the microbiomes of Tamworth and feral pigs: both greatly underrepresented in the literature.

Currently there are an estimated 6 million feral pigs in the United States [6]. Feral pigs were first introduced in the early 1500s by Spanish settlers and cause significant ecological damage. It has been shown that feral pigs decrease the amount of plant litter and cover in areas they feed [7]. Yet, with the ecological and economic toll feral pigs exert, little study has been conducted to elucidate the structure of their microbiome. The Tamworth breed is thought to be descended from the Old English Forest pig and has not

been crossed or improved with other breeds since the late 18th Century [British Pig 8]. The breed is not a traditional animal used in high production agriculture, bred instead for its tolerance to cold weather and ability to forage. Additionally, the Tamworth breed is under watch by the Livestock Conservancy, after previously being designated as threatened, and the microbiome composition has yet to be characterized.

In this study, we characterized the microbiomes of Tamworth and feral pigs using a combination of culture-dependent and independent techniques on direct colon and cecum contents. To date, modern high throughput culture efforts have been reserved almost exclusively to human fecal samples. Here we extend such methodology to pigs. The culture strategy employs a single medium, yBHI, with various selection screens to shift taxa retrieval. A single medium isolation strategy will facilitate downstream defined community studies. For example, simple to complex bacterial communities can be assembled in bioreactors to study the mechanisms of pig gut microbiome succession [9]. Similarly, colonization of such defined communities constituted could reveal how gut bacterial species or combinations impact gut development and immunity [10]. Availability of a well characterized strain library with genome information will facilitate future studies to better understand the role of pig gut microbiome in health and disease.

MATERIALS AND METHODS

Sample Collection and Preparation

Permission was granted from purchasers of three Tamworth pigs to obtain colon and cecum samples immediately following slaughter. The Tamworth pigs sampled here were not given any antibiotics or growth promoters and could freely graze. Small incisions were made into either the colon or cecum with a sterile disposable scalpel. Lumen

contents were gently squeezed into sterile 50 mL tubes, mixed with an equal proportion of 40% anaerobic glycerol (final concentration 20% anaerobic glycerol), and immediately snap frozen in liquid nitrogen. For culture preparation, samples were pooled under anaerobic conditions in a vinyl chamber (Coy Labs, USA). Feral samples were kindly provided by boar hunters in Texas, US. A similar procedure was followed where colon and cecum samples were taken immediately following evisceration, mixed with anaerobic glycerol and frozen.

Metagenomics

DNA was extracted from gut samples using the DNeasy PowerSoil kit (Qiagen, Germany) following the provided kit protocol. After extraction, Microbial DNA was enriched with the NEBNext® Microbiome DNA Enrichment Kit (New England Biolabs, US) to remove host DNA present after DNA extraction. Metagenomic sequencing was conducted on the Illumina MiSeq platform utilizing V2 (250 bp) paired-end sequencing chemistry. Raw sequencing reads were quality controlled using the read-qc module in the software pipeline metaWRAP [11]. Briefly, reads are trimmed to PHRED score of > 20 and host reads not removed by enrichment were removed by read-mapping against a reference pig genome (GCF_000003025.6). Resultant reads from read-qc are hereby referred to as high-quality reads. High-quality reads were passed to Kaiju [12] for taxonomy annotation against the proGenomes database (http://progenomes.embl.de/, downloaded March 1, 2019). Kaiju was run in default greedy mode and resultant annotation files were parsed in R [13]. Mash [14] was run to estimate the Jaccard distance between samples. 10,000 sketches were generated for each sample and the sketches were compared using the *dist* function provided in the Mash software.

Antimicrobial resistance (AMR) genes were predicted from metagenomics assemblies. High-quality sequencing reads were assembled into contigs using the assembly module in metaWRAP ; metaSPAdes [15] was the chosen to assemble the reads: contigs greater than 1,000 bp were retained. Prodigal [16] was run to predict open reading frames (ORF) using the metagenomic training set. Abricate [17] was then run to annotate the ORF against the NCBI Bacterial Antimicrobial Resistance Reference Gene Database (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA313047, downloaded April 22, 2019).

Contigs were gathered into bins using three methods: MetaBAT2 [18], MaxBin2 [19], and CONCOCT [20]. Contig bins were kept if the contamination was less than 5% and bin completeness was greater than 85% as determined by CheckM [21]. Bins from the three methods used were refined into a coherent bin set using the bin_refinement module in metaWRAP. Refined bins were reassembled with a minimum contig length of 200 bp and the same contamination and completeness parameters as initial bin construction. Metagenomic bin and pure isolate phylogeny was generated using UBCG [22] to identify and align 92 marker genes. Tree construction was conducted using RAxML [23] : GTR+G4 nucleotide model. To identify KEGG homologues, ORF were identified in metagenomic assemblies, bins, and culture genomes using Prodigal. The resultant ORF were annotated against the KEGG database using KofamKOALA [24] run locally.

Culturomics

Colon and cecum samples were pooled respective to feral and Tamworth samples before culture experiments. All culture experiments, including pooling, were conducted under

anaerobic conditions inside an anaerobic chamber (Coy Labs, USA). Samples were serially diluted in sterile anaerobic PBS and spread plated onto the media conditions listed in supplemental table 1. Plates were inoculated at 37°C for 48 hours before initial colony selection. 25 colonies were non-selectively sub-cultured from the initial plate to yBHI plates. The procedure was repeated after 72 hours for a total of 50 colonies per media condition. Colonies were primarily identified using MALDI-TOF (Bruker, Germany). MALDI-TOF scores greater than 2.0 were considered a positive species identification. Scores between 1.7 - 2.0 were taken as positive genus identification. Colonies without a positive MALDI-TOF identification were identified by sequencing the 16s rRNA gene. Briefly, DNA was extracted from colonies using the DNeasy Blood and Tissue Kit (Qiagen, Germany) following the manufacturer's protocol. 16s rRNA sequence was amplified using 27F and 805R primers. The primer sequence is listed in supplemental table 1. Genomes of the selected strains were sequenced on the MiSeq platform utilizing paired-end v3 chemistry (300 bp). Sequencing reads from individual strains were assembled with Unicycler [25] : minimum contig length of 200 bp. The raw sequencing reads from the culture isolates and metagenomic samples are hosted at NCBI under the BioProject ID PRJNA555322.

RESULTS

Metagenomics describes community composition

The gut microbiota of Tamworth breed and feral pigs are underrepresented in the literature. As such, the pigs provide a unique model to study the implementation of culture-dependent and independent techniques. To begin the investigation, colon and cecum samples were metagenomically sequenced from both breeds. Figure one shows

the taxonomic annotation of the metagenomic reads respective to the source of isolation (Fig 1(A)(B)). Shotgun metagenomic sequencing provides an accurate description of the community outside the bias of culture-dependent methods. By incorporating sequencing, the efficiency of the culture method can be elucidated. The phylum Bacteroidetes represents nearly 53% of all classified reads in Tamworth pigs, compared to 29% in feral pigs. The abundance of Firmicutes in Tamworth samples is lower than feral samples at 15% and 28% respectively. Additionally, nearly 10% more of the feral reads were unclassified compared to Tamworth (37%, 28%). Turning to the genus level, the large increase of Bacteroidetes in Tamworth pigs is primarily composed of the genus *Prevotella*, Figure 1 (B) (38%, feral 11%). Remarkably, the genus *Bacteroides* showed almost identical distribution between the feral and Tamworth pigs (7.6% and 7.6% respectively). The increase of Firmicutes in feral samples is due to an increase in several genera such as *Ruminococcus*, *Clostridium*, and *Eubacterium.* This corresponds with significantly higher Shannon diversity index values in feral samples compared to Tamworth ($p = 0.0024$, Wilcoxon rank-sum test). Full phylum and genus annotation tables are provided in supplemental table 2.

Another advantage of implementing culture-independent methods is the speed and ease with which samples can be clustered. To cluster the sequencing results from the microbiome samples, Mash [14] was used to sketch the reads sets and compile a distance matrix (Figure 1C). Within the matrix, both Kmeans clustering and hierarchical clustering (average-linkage) readily separated the input samples according to isolation source. Mash provides a method to compare metagenomes that is not subject to annotation bias. Principal component analysis (PCA) of the OTU tables was the second

method employed to cluster the metagenome results. Again, two distinct groups corresponding to Tamworth and feral samples are seen in the plot (figure 1D). Interestingly, the Tamworth samples are more homogenous in both the Mash and PCA methods. All pigs were taken from the same farm which may account for the lower inter-animal microbiome divergence. The herd status of the feral pigs is unknown.

With metagenomic sequencing, the AMR homologues from a bacterial community can be identified. Briefly, the sequencing reads are assembled into contigs, open reading frames are predicted, and annotated against relevant databases of AMR genes. AMR homologues were identified in the gut samples as the culturing scheme depends on antibiotic selection to shift bacterial diversity. High loads of antibiotic resistant bacteria in fecal samples would inhibit the culture scheme. The gut samples from Tamworth pigs' all yielded at least eight AMR homologues (Figure 2). Additionally, four AMR genes: *cfxA, lnu(AN2), mef(En2), and tet(40)* were identified in every Tamworth sample. In general, the level of AMR genes in Tamworth samples was higher than feral samples. No common pattern is apparent for Feral samples; however, *tet(Q)* is found in 5 of 9 feral samples. The full result of the AMR gene query is listed in supplemental table 3.

Selective screens shift plating diversity

High throughput anaerobic culturing was the second method employed to study the microbiota of Tamworth and feral pigs. We implemented culture-dependent methods, in addition to sequencing methods, as we believe the two will provide a more coherent view of a microbiomes structure and function. The culture sampling strategy utilized is as follows: a base medium (yBHI, or close derivatives) had various selective screens

(antibiotics, heat, bile, etc.,) applied to it. A finite growing surface is available for colonization and some species will grow more rapidly and subsequently outcompete others. If appropriate selective pressure is applied, we hypothesized that interspecies selection would decrease allowing for taxa not retrieved in plain medium conditions to grow. The approach is similar to one previously used to culture strains from human fecal samples [26]. One major difference is said work used multiple media compositions, rather than one as in our study. Ten media conditions were used for both Tamworth and feral samples and are listed in supplemental table 1. 25 colonies were picked at 48- and 72-hours post inoculation, for a total of 50 colonies per condition. In total, 1000 colonies were selected from plates, of which 884 were successfully identified. Selective screens shifted the taxa retrieved (figures 3). Figure 3 depicts the number of isolates per media condition with a bar plot depicting the total number of isolates retrieved. *Lactobacillus sp*. was the most abundant organism retrieved (166 isolates) followed by *Escherichia coli* (86), *Lactobacillus mucosae* (74) and *Streptococcus hyointestinalis* (64). The top ten isolates cultured are listed in table 1. One case of selection completely changing plate diversity compared to plain media is that of heat shock treatment. As expected, many spore forming genera including *Bacillus* and *Clostridium* were only able to grow when the inoculum was heated to kill vegetative cells. The selective screens placed upon yBHI not only shifted the taxa retrieved from each plating condition as shown in Figure 3, but also shifted species richness and evenness (figure 4). The most diverse plating condition (Shannon Index) for both Tamworth and feral samples was obtained from plain yBHI: showing as a log-normal community distribution. Similar log-normal community structures are observed for BSM (Tamworth only), Erythromycin and heat shock

treatments. Bile treatments and chlortetracycline exhibited strong selective pressure shown as geometric series in the species-rank abundance plots (figure 4). Most of the taxa retrieved from the bile condition were identified as Proteobacteria, indicating that the dosage of bile (1 g / L) was too high.

The culture strategy did not recapitulate the community in the inoculum as defined by metagenomics. In both Tamworth and feral samples, a high number of Firmicutes and Proteobacteria were isolated, compared to the metagenomic sampling where Bacteroidetes was the most abundant phylum for both sources. If we disregard the bile conditions, which were dominated by Proteobacteria, yBHI clearly selects for common Firmicutes genera including: *Lactobacillus, Streptococcus,* and *Bacillus.* While the screens were successful in increasing the total number of species retrieved, no condition matched the inoculum in form. *Prevotella* for example, the most abundant genus in Tamworth pigs, was only retrieved seven times from 500 colonies. Taken together, the strategy was successful in gathering many isolates that can grow on a common medium but failed in that the most abundant taxa were not retrieved in proportion to the inoculum.

Culturing captures genomic information not captured in metagenomics

The sampling strategy employed did not recapitulate the inoculum community. However, one of the main reasons we chose to culture was that we believed rare taxa would provide information that would be loss to metagenomics. To examine this, we sequenced selected isolates and generated 81 high quality metagenomic bins (completeness > 85%, contamination < 5%). The phylogeny of the metagenomic bins and culture genomes was estimated (figure 5). Consistent with read taxonomy, many of the bins constructed from

both Tamworth and feral samples were annotated to the phylum Bacteroidetes. The phyla Firmicutes, Proteobacteria and Actinobacteria were comprised almost entirely of isolate genomes. Isolate genomes not only populated clades of the tree missed by metagenomic bins, but provided genes not observed in metagenomic assemblies nor bins (figure 6). Open reading frames (ORF) were predicted from metagenomic assemblies, metagenomic bins, and culture isolate and were annotated against the KEGG database. Figure 6 shows the abundance (natural log) of KEGG homologues respective to the source of the ORF. The full KEGG annotations from the bins, isolates, and metagenomes are provided in supplemental table 4. Metagenomic bins contained less information than the metagenomic assemblies. This is expected as the bins are derived from contigs in the assemblies and not all the contigs will be gathered into bins. The isolates however provided KEGG homologues that were completely missed through culture-independent methods. Thus, culture and culture-independent methods can augment a microbiota analysis providing information that the other method cannot capture.

DISCUSSION

In this work, we establish a methodology that provides a rigorous examination of the pig gut microbiome. Instead of viewing culture-dependent and independent methods as a binary decision, we contend that both techniques complement one another. Community composition and gene content is readily identified by metagenomics. In addition, metagenomic techniques require less labor than culture-dependent ones. As a result, current research on the pig gut microbiota has almost exclusively focused on culture-independent methods [27-29]. Metagenomic sequencing however cannot establish the role specific taxa play in microbial communities. Recent works on the gut

microbiome of humans have shown the value of culture-dependent techniques, especially in identifying mechanisms that govern bacterial communities [30, 31]. Incorporation of culture-techniques would greatly increase our understanding of how microbial communities form and persist in pigs.

Despite the high abundance of *Prevotella* in both Tamworth and feral pigs, the culture sampling strategy we employed only yielded seven *Prevotella* isolates from Tamworth samples (7/500, 1.4%). No *Prevotella* was isolated from the feral inoculum. In contrast, several genera including *Lactobacillus*, *Escherchia*, *Streptococcous*, and *Bifidobacterium* were overrepresented in culture samples as compared to metagenomic sequencing. Our culture results align with an early culture examination of the pig microbiome. In that study, the two most abundant isolates cultured were gram-positive cocci and *Lactobacillus* [32]. Both our work and the earlier work relied upon complex media derived largely of peptone digests. As *Prevotella* is associated with an increase of dietary fiber, work will be needed to develop a defined media that is not based upon peptides such as yBHI. Previous studies have been wildly successful in culturing many bacteria that were previously thought to be “unculturable” [33, 34]. In those works, samples were plated onto multiple media formulations to encourage a broad growth of bacterial species. While the multiple media approach generates a higher number of taxa, one study isolated over 1,300 species [33], creating multiple media formulations is expensive and time-consuming. Additionally, bacteria isolated from different media formulations may not grow together on a common media, forfeiting any combined in vitro experimentation. The approach we implemented ensures all isolates can grow on a common media source.

Recent studies have proposed metagenomic binning as a culture-independent method to extract genomes from samples [28, 35-37]. However, one of the main pitfalls of metagenomic binning is that metagenomic assemblers struggle to assemble contigs of closely related taxa, especially if the organisms are found in low abundance [38]. With knowledge now that strain-level variation occurs in species of the microbiome [39], targeted culture efforts are needed to confirm that strain variation observed in metagenomic data is not simply due to assembler bias. Also, culture isolates provided genetic information that was not captured in the metagenomic sequencing.

ACKNOWLEDGEMENTS

Computations supporting this project were performed on High-Performance Computing systems managed by Research Computing Group, part of the Division of Technology and Security at South Dakota State University. This work was in part supported by the grants from the South Dakota Governors Office of Economic Development (SD-GOED) and the United States Department of Agriculture (grant numbers SD00H532-14 and SD00R646-18) awarded to JS. We gratefully thank Tom Harrison, Jeff Hopper and Brett Grogan for their help in collecting feral pig gut microbiota samples. We are saddened by the untimely passing away of Tom Harrison during the course of this study. We dedicate this manuscript to Tom’s memory.

Figures

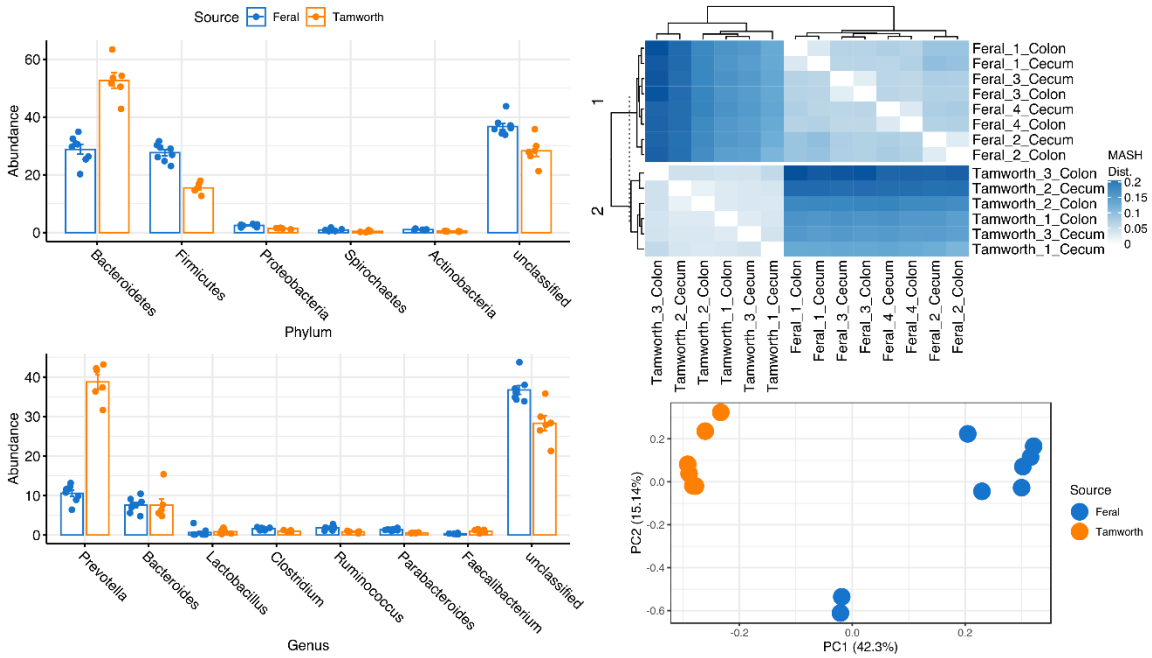


Fig 1. Metagenomic analysis of feral and Tamworth colon and cecum samples.(A)(B) Relative abundance of major phyla and genera annotated from sequencing reads respective to isolation source. (C) Matrix depicting the MASH distance between feral and Tamworth samples. Clusters 1 and 2 are defined by kmeans clustering. (D) Principal component analysis of the taxon abundance obtained from the read annotation. Samples are colored respective to isolation source.

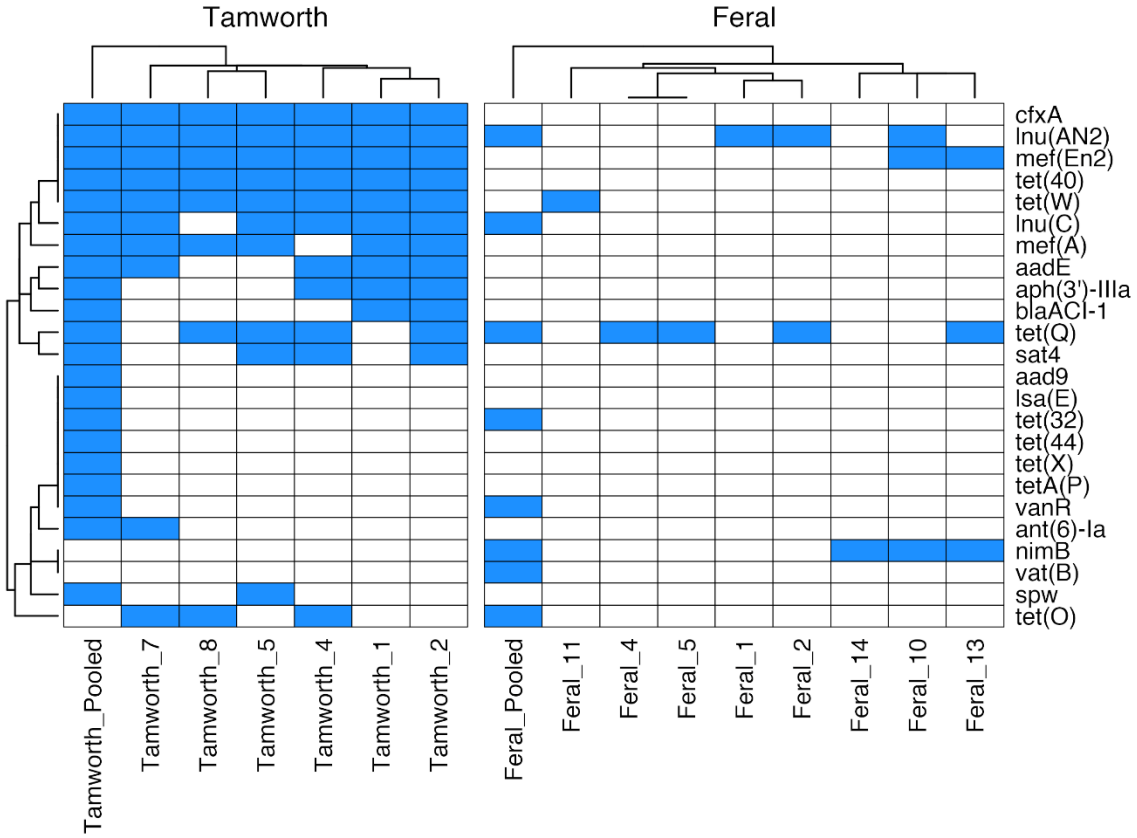


Fig 2. Antimicrobial resistance (AMR) homologues annotated from metagenomic samples. Columns depict individual samples and rows correspond to AMR homologues. Blue color depicts the presence and white color corresponds to absence. AMR homologues were considered present if the coverage value was greater than 90% and a percent homology greater than 70%.

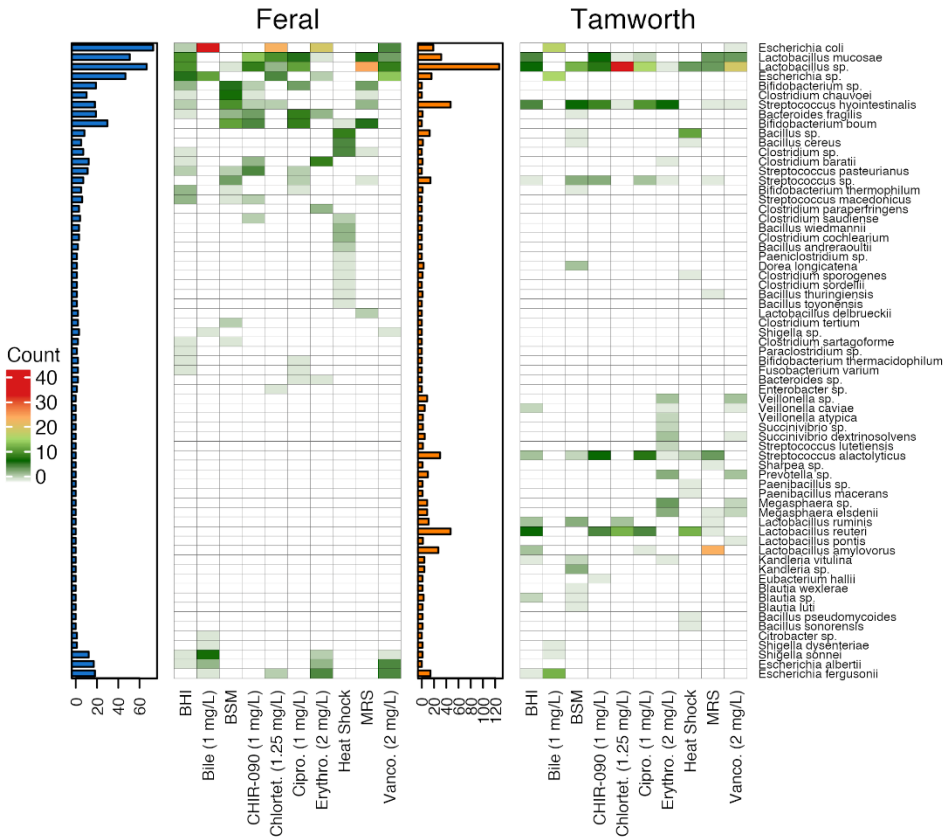


Fig 3. Bacteria isolated from various media conditions. Columns represent individual media conditions and row correspond to bacterial taxa retrieved, cells are colored respective to the number of isolates cultured per media condition. The corresponding bar plot to the left of the matrices shows the total number of isolates retrieved per isolation source.

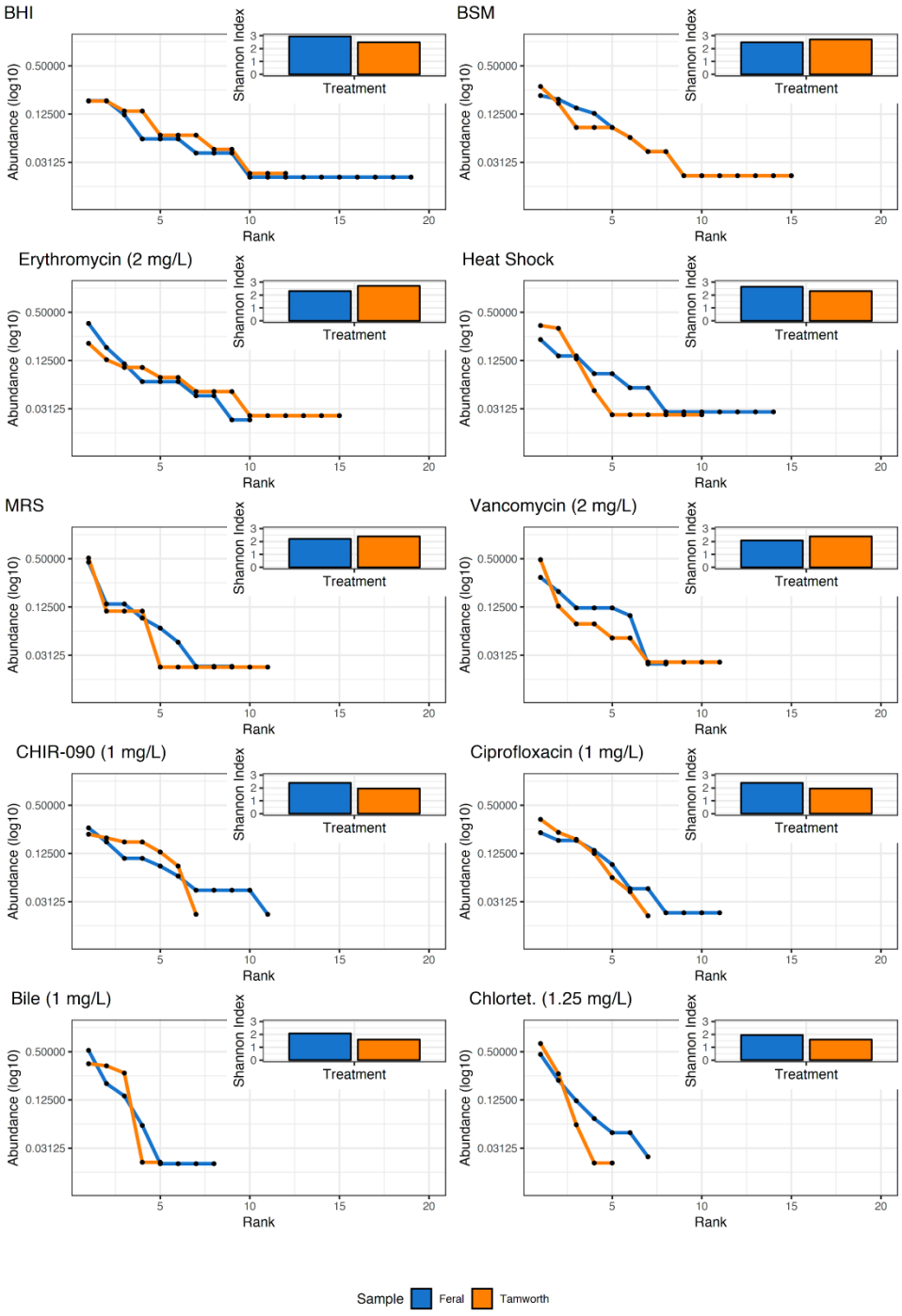


Fig 4. Rank abundance curves of the various media conditions. The community evenness of the various media conditions is shown respective to the isolation source. The inlay plot depicts the Shannon Index respective to the isolation source.

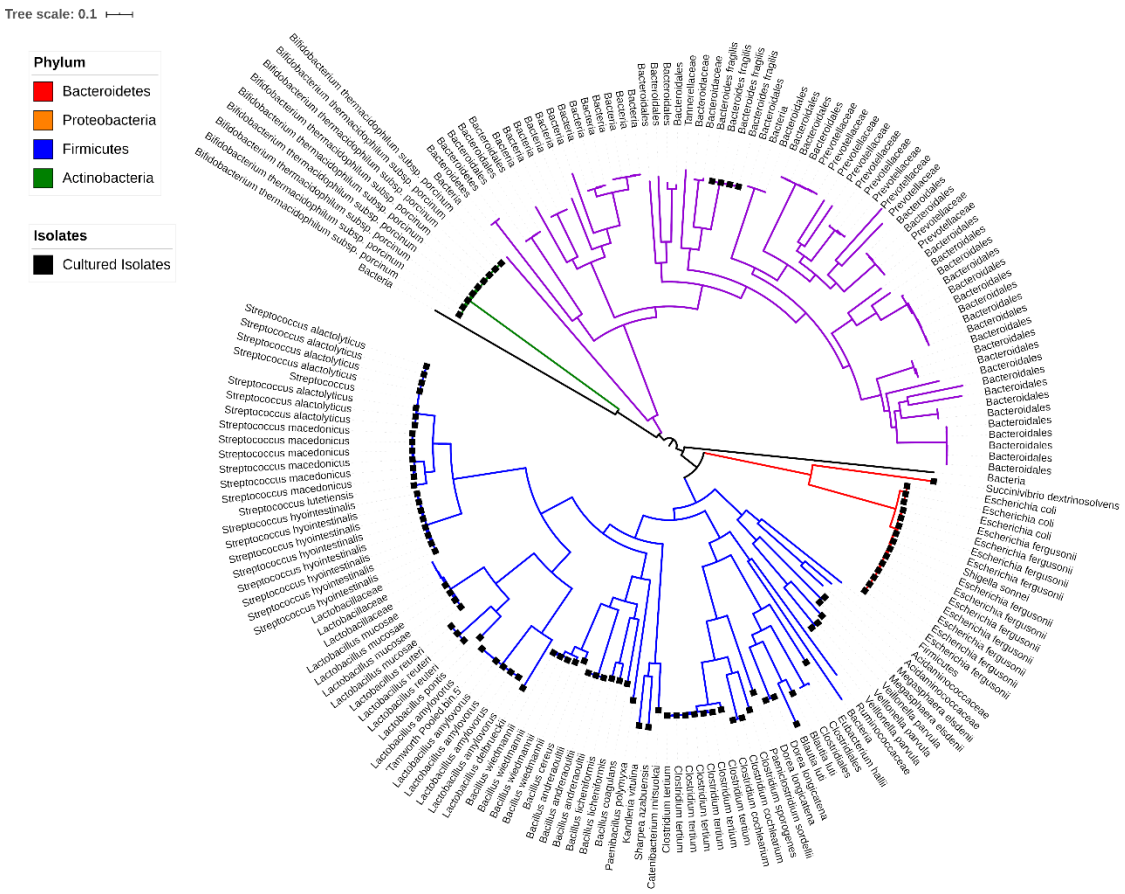


Fig 5. Maximum-likelihood tree of metagenomic bins and culture genomes.Tree was constructed from a nucleotide alignment of 92 single-marker genes. General time reversible (GTR) was chosen as the substitution model in tree construction.

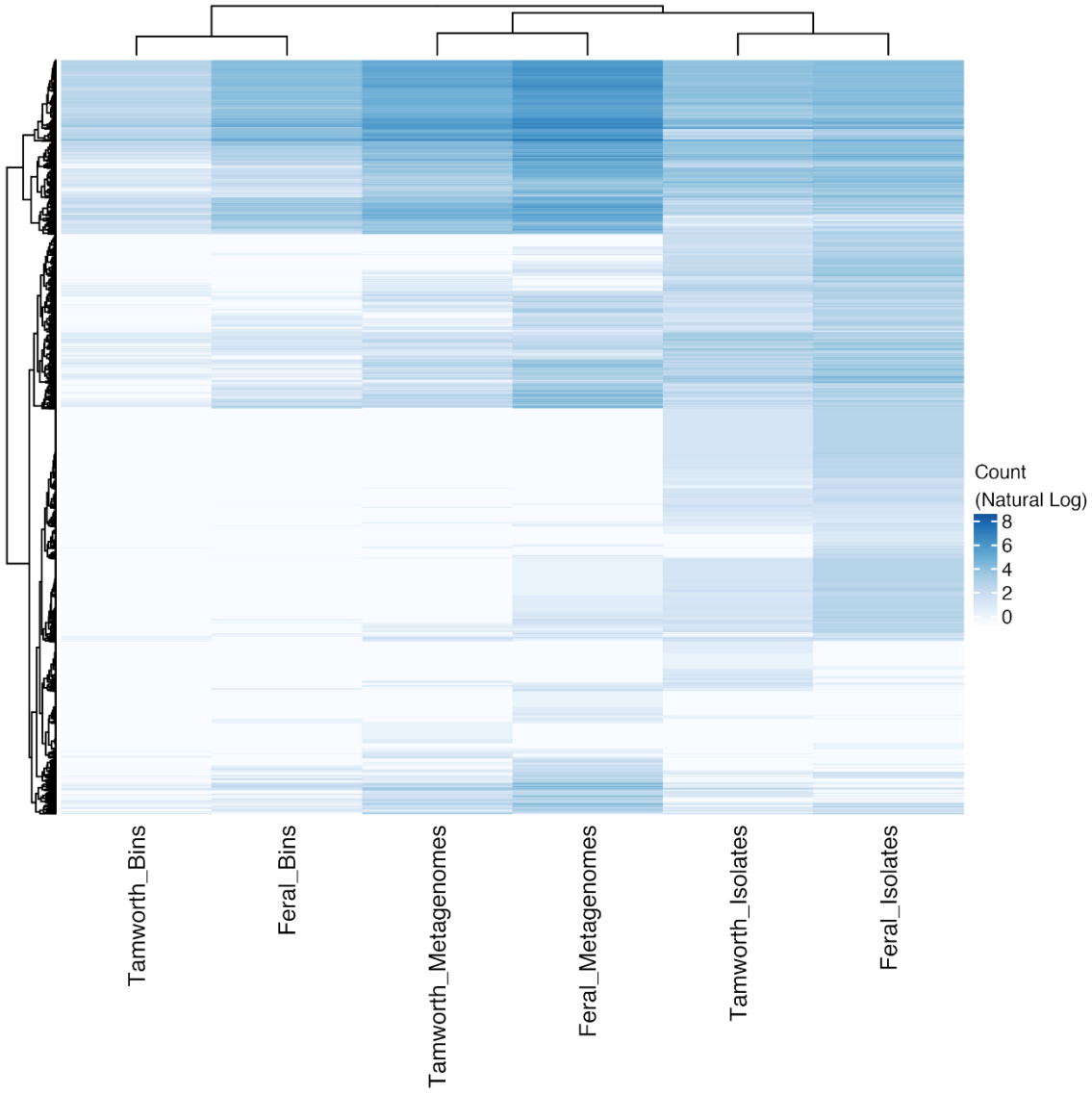


Fig 6. KEGG annotation of open frames from metagenomic assemblies, bins, and culture genomes. Rows and columns are clustered using an average linkage method. KEGG annotations counts are represented as the natural log to increase the clarity of the figure.

REFERENCES

1. Singh, K.M., et al., *Taxonomic and gene-centric metagenomics of the fecal microbiome of low and high feed conversion ratio (FCR) broilers.* Journal of Applied Genetics, 2014. **55**(1): p. 145-154.
2. Piewngam, P., et al., *Pathogen elimination by probiotic Bacillus via signalling interference.* Nature, 2018. **562**(7728): p. 532-537.
3. Byndloss, M.X., et al., *Microbiota-activated PPAR-gamma signaling inhibits dysbiotic Enterobacteriaceae expansion.* Science, 2017. **357**(6351): p. 570-575.
4. Hedegaard, C.J., et al., *Natural Pig Plasma Immunoglobulins Have Anti-Bacterial Effects: Potential for Use as Feed Supplement for Treatment of Intestinal Infections in Pigs.* PLOS ONE, 2016. **11**(1): p. e0147373.
5. Metzler-Zebeli, B.U., et al., *Adaptation of the Cecal Bacterial Microbiome of Growing Pigs in Response to Resistant Starch Type 4.* Applied and Environmental Microbiology, 2015. **81**(24): p. 8489.
6. USDA. *History of Feral Swine in the Americas*. 2018; Available from: https://www.aphis.usda.gov/aphis/ourfocus/wildlifedamage/operational-activities/feral-swine/sa-fs-history.
7. Siemann, E., et al., *Experimental test of the impacts of feral hogs on forest dynamics and processes in the southeastern US.* Forest Ecology and Management, 2009. **258**(5): p. 546-553.
8. Association, B.P. *The Tamworth*. n.d.; Available from: http://www.britishpigs.org.uk/breed_tw.htm.
9. Auchtung, J.M., C.D. Robinson, and R.A. Britton, *Cultivation of stable, reproducible microbial communities from different fecal donors using minibioreactor arrays (MBRAs).* Microbiome, 2015. **3**: p. 42.
10. Goodman, A.L., et al., *Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice.* Proc Natl Acad Sci U S A, 2011. **108**(15): p. 6252-7.
11. Uritskiy, G.V., J. DiRuggiero, and J. Taylor, *MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis.* Microbiome, 2018. **6**(1): p. 158.
12. Menzel, P., K.L. Ng, and A. Krogh, *Fast and sensitive taxonomic classification for metagenomics with Kaiju.* Nature Communications, 2016. **7**: p. 11257.
13. R Core Team, *R: A Language and Environment for Statistical Computing*. 2019.
14. Ondov, B.D., et al., *Mash: fast genome and metagenome distance estimation using MinHash.* Genome Biology, 2016. **17**(1): p. 132.
15. Nurk, S., et al., *metaSPAdes: a new versatile metagenomic assembler.* Genome Res, 2017. **27**(5): p. 824-834.
16. Hyatt, D., et al., *Prodigal: prokaryotic gene recognition and translation initiation site identification.* BMC bioinformatics, 2010. **11**: p. 119-119.
17. Seemann, T., *ABRicate*. 2018: GitHub.
18. Kang, D., et al., *MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies.* PeerJ Preprints, 2019. **7**: p. e27522v1.

19. Wu, Y.W., B.A. Simmons, and S.W. Singer, *MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets.* Bioinformatics, 2016. **32**(4): p. 605-7.
20. Alneberg, J., et al., *Binning metagenomic contigs by coverage and composition.* Nature Methods, 2014. **11**: p. 1144.
21. Parks, D.H., et al., *CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.* Genome research, 2015. **25**(7): p. 1043-1055.
22. Na, S.-I., et al., *UBCG: Up-to-date bacterial core gene set and pipeline for phylogenomic tree reconstruction.* Journal of Microbiology, 2018. **56**(4): p. 280-285.
23. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics (Oxford, England), 2014. **30**(9): p. 1312-1313.
24. Aramaki, T., et al., *KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold.* bioRxiv, 2019: p. 602110.
25. Wick, R.R., et al., *Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads.* PLOS Computational Biology, 2017. **13**(6): p. e1005595.
26. Rettedal, E.A., H. Gumpert, and M.O.A. Sommer, *Cultivation-based multiplex phenotyping of human gut microbiota allows targeted recovery of previously uncultured bacteria.* Nature Communications, 2014. **5**: p. 4714.
27. Xiao, L., et al., *A reference gene catalogue of the pig gut microbiome.* Nature Microbiology, 2016. **1**(12): p. 16161.
28. Wang, W., et al., *Metagenomic reconstructions of gut microbial metabolism in weanling pigs.* Microbiome, 2019. **7**(1): p. 48.
29. Tan, Z., et al., *Metagenomic Analysis of Cecal Microbiome Identified Microbiota and Functional Capacities Associated with Feed Efficiency in Landrace Finishing Pigs.* Frontiers in Microbiology, 2017. **8**(1546).
30. Poyet, M., et al., *A library of human gut bacterial isolates paired with longitudinal multiomics data enables mechanistic microbiome research.* Nature Medicine, 2019. **25**(9): p. 1442-1452.
31. Gutiérrez, N. and D. Garrido, *Species Deletions from Microbiome Consortia Reveal Key Metabolic Interactions between Gut Microbes.* mSystems, 2019. **4**(4): p. e00185-19.
32. Russell, E.G., *Types and Distribution of Anaerobic Bacteria in the Large Intestine of Pigs.* Applied and Environmental Microbiology, 1978. **37**(2): p. 187-193.
33. Lagier, J.-C., et al., *Culture of previously uncultured members of the human gut microbiota by culturomics.* Nature Microbiology, 2016. **1**: p. 16203.
34. Browne, H.P., et al., *Culturing of 'unculturable' human microbiota reveals novel taxa and extensive sporulation.* Nature, 2016. **533**: p. 543.
35. Albertsen, M., et al., *Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes.* Nature Biotechnology, 2013. **31**: p. 533.

36. Tully, B.J., E.D. Graham, and J.F. Heidelberg, *The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans.* Scientific Data, 2018. **5**: p. 170203.
37. Pasolli, E., et al., *Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle.* Cell, 2019. **176**(3): p. 649-662.e20.
38. Ayling, M., M.D. Clark, and R.M. Leggett, *New approaches for metagenome assembly with short reads.* Briefings in Bioinformatics, 2019.
39. Lloyd-Price, J., et al., *Strains, functions and dynamics in the expanded Human Microbiome Project.* Nature, 2017. **550**(7674): p. 61-66.

CHAPTER 2: Geography Shapes the Population Genomics of *Salmonella enterica* Dublin

INTRODUCTION

Salmonella enterica serotype Dublin (*S.* Dublin) is a host-adapted serotype of *Salmonella enterica* that is primarily associated with cattle. In contrast to many enteric diseases affecting cattle that are presented primarily with diarrhea in young calves, *S.* Dublin infection can manifest as both enteric and systemic forms in older calves (1). When cattle ingest sufficient infectious dose of *S.* Dublin, typically greater than 10^6 CFU's (2), *S.* Dublin could colonize the gut of the animal. After colonization, *S.* Dublin invades enteric cells in the ileum and jejunum and subsequently traverses to the mesenteric lymph nodes ultimately causing systemic infection (3). It has been shown that a virulence plasmid carried by *S.* Dublin is partly responsible for the systemic phase of the infection; removal of the plasmid or the Salmonella plasmid virulence (spv) genes carried upon the plasmid attenuates systemic infections (1, 4). Analogous to Salmonella Typhi infections in humans, *S.* Dublin is known to establish a carrier state in susceptible cattle. Carrier animals harbor the bacteria in internal organs and lymph areas and sporadically shed *S.* Dublin through feces and milk (5). Such carriers tend to help to maintain *S.* Dublin infection rates in local dairy herds and cases of human infections after drinking raw milk contaminated with the pathogen have been documented (6-8). The duration and severity of shedding is highly variable between animals. Some animals may begin shedding *S.* Dublin in feces as soon as 12 – 48 hours after infection (2). Shedding has been detected up to six months after the initial discovery that an animal is a carrier (5).

Current genomics of *S.* Dublin has primarily focused upon the identification of antimicrobial resistance (AMR) homologues and mobile genetic elements such as prophages and plasmids (9-11). *S.* Dublin genome diversification appears to be driven by horizontal gene transfer and genome degradation resulting in pseudogenes (11). However, many of these studies focused on a smaller set of *S.* Dublin, typically less than 30 genomes, and focused on comparisons to closely related serotypes. The population structure of *S.* Dublin has yet to be resolved, especially regarding isolates from outside of the US. Due to the importance of the pathogen in animal agriculture and human health, establishing the population structure and pangenome of the serotype would provide valuable insight into the evolution of *S.* Dublin. Phylogeographical clustering is evident in the population of *S.* Dublin and impacts the composition of the core and ancillary genomes.

MATERIALS AND METHODS

Genome sequencing and comparative data set. We sequenced 43 *S.* Dublin clinical isolates that were collected by the Animal Disease Research Laboratory (ADRDL, South Dakota State University) and 33 isolates collected by the Animal Health Diagnostic Center (AHDC, Cornell University). Strains were grown aerobically in Luria Bertani broth at 37°C for 12 hours. DNA was isolated from resultant pellets using the DNeasy Blood and Tissue Kit (*Qiagen,* Hilden, Germany). Paired-end sequencing was conducted using the Illumina MiSeq platform and 250 base paired V2 chemistry. For the comparative genome analysis and the construction of global population structure, 1020 publicly available *S.* Dublin genome data was downloaded from the NCBI Sequence Read Archive (SRA). Raw sequence data as well as the metadata tables were

downloaded and manually parsed to include samples that contained a positive *S.* Dublin serotype that were sequenced using the Illumina platform. The prefetch utility (SRA toolkit) was used to download the SRA files which were written into paired-end fastq files with the fastq-dump tool (SRA toolkit).

Genome assembly and validation. Paired-end reads were assembled into contigs using Shovill (33) given the following parameters : minimum contig length 200, depth reduction 100x. and an estimated genome size of 4.8 Mbp. The Shovill pipeline is as follows: read depth reduction per sample to approximately 100x of the estimated genome size; read sets below the 100x threshold are not affected by the reduction. After reduction, reads are conservatively error corrected with Lighter (34). Spades (v3.12.0) (35) was used to generate the assembly using default parameters. After assembly, small indels and assembly errors are corrected using Pilon (36). Genome assemblies were passed to the software assembly-stats (https://github.com/sanger-pathogens/assembly-stats) to gauge basic assembly properties such as contig number, N50, and genome length. Samples were eliminated from the data set if the assemblies were fragmented, defined here as a contig number greater than 300 or an N50 less than 25,000 bp.

Serotype prediction. Genomes that passed assembly validation were submitted to serotype validation. The program SISTR (12) was download and run locally to validate the serotypes using both cgMLST, SISTR method, and Mash (13). A positive Dublin serotype was defined as a confirmed Dublin prediction from both the Mash and cgMLST identification methods.

Pangenome Reconstruction. Curated genome assemblies were annotated using the software Prokka (37). A manually annotated reference *Salmonella enterica* Typhimurium

(ASM694v2) genbank file was downloaded(ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/Salmonella_enterica/reference/GCF_000006945.2_ASM694v2) and formatted to a Prokka database file. Said reference database was used to augment the existing Prokka databases and facilitated consistent nomenclature of core Salmonella genes. Resultant general feature files 3 (.gff) from Prokka were used as the input to the program Roary (38). PRANK (39) was used within Roary to conduct the alignment of core genes.

Phylogeny reconstruction. Two distinct methods were used to generate phylogenomic trees. In the first method, the core gene alignment file from Roary was passed to the software SNP-Sites (40) using the flags –cb to discard gaps and include monomorphic sites. The final alignment file is 2,431,413 base pairs longs. Model-test NG (https://github.com/ddarriba/modeltest) was used to define a substitution model for phylogeny and was run to optimize a model for RAxML. Generalized Time Reversible (GTR) + G4 was the best scoring model and used to generate the maximum-likelihood tree. The interactive Tree of Life (iTOL) (41) was used to visualize phylogenomic trees. kSNP3 (42) was the second method employed to generate a reference independent SNP phylogeny based upon kmers, rather than gene polymorphisms. The program was run to generate a fasta matrix based upon kmers found in 99% of genomes. Said matrix was passed to Model-test NG same as above and a maximum-likelihood tree was generated using RAxML (43) with the GTR model.

BEAST2 Phylogeny. The core gene alignment file from Roary was passed to SNP-Sites given the flags -cb to generate an alignment for BEAST2 (14). BEAUti was used to generate the xml using a strict molecular clock and a constant coalescent population

model. GTR+G4 was used for the nucleotide substitution model. The chain length was set to 10,000,000 sampling every 1000 trees for the log file with a seed value of 33.

Antibiotic resistance homolog prediction. Antibiotic resistance gene homologs were predicted in the genomes using the software package Abricate (44). The NCBI Bacterial Antimicrobial Resistance Reference Gene Database was used as a reference. Positive hits are defined here as homologs with greater than 90% sequence identity and greater than 60% of target coverage.

Plasmid replicon identification. Analogues to AMR homologue detection. Abricate was used to BLAST genomes assemblies against the PlasmidFinder (45) database for identification of plasmid replicons. Positive hits were defined as a sequence identity > 90% and at least 60% query coverage.

Identification of prophage regions. The web service PHASTER was used to identify prophage regions in the genomes (13, 14). As the number of genomes in the query was high, a bash script was used to submit genome assemblies via the API provided. Resulting text files from PHASTER were downloaded, concatenated, and parsed using R (R Core Team, 2018). Prophage regions were considered if they were marked as “intact”.

Data analysis. Logistic PCA was conducted using the R package logisticPCA (17). Binary, 0 and 1, presence-absence matrices were prepared from the Roary output with the rows corresponding to genes and the columns corresponding to genomes. Genes with a prevalence greater than 99% or less than 5% were removed to aid in computational speed and reduced confounding effects of misannotation. All other statistical analysis and plotting were conducted using R using the packages: ggplot2(46), ggtree (47), and ComplexHeatmap(48).

RESULTS

S. Dublin global population structure. To begin the investigation on *S.* Dublin, 74 isolates of *S.* Dublin were sequenced using the Illumina MiSeq platform. For comparative analysis, 1020 publicly available *S.* Dublin genomes were downloaded from NCBI Sequence Read Archive (NCBI SRA). Genomes were assembled and subjected to a two-step validation process. The first validation step was to assess genome assembly quality; genome assemblies with greater than 300 contigs or a N50 values less than 25,000 base pairs were discarded from the analysis set. The second validation step was serotype verification using the program SISTR (12). Genomes were retained in the dataset if the serotype prediction using core genome multilocus sequence typing (cgMLST) and MASH (13) agreed on a prediction of *S.* Dublin. After assembly and validation, a high-quality dataset of 880 genomes were used for further analysis. Table 1 describes the full dataset and the full metadata for the genome dataset is provided in supplemental table 1. Based on their origin, genomes are grouped into 4 major geographical regions: Africa (4%), Brazil (13%), the United Kingdom (20%), and the United States (62%). In terms of host species of isolation, human genomes are the largest constituent representing nearly 38% of genomes followed by bovine isolates (30%), food isolates (24%) and isolates from various sources or without metadata were classified as other (8%). Clear sampling bias is evident in the publicly available genome dataset as more than half of the genomes retrieved are of US origin. Additionally, nearly 88% (231/262) of bovine isolates originate from the US.

Core genome variation was first investigated using core gene (n=4,098) polymorphism trees (figure 1, 2, supplemental figure 1). Core genome phylogeny revealed strong geographical demarcation between genomes (figure 1). Five major clades are seen in the phylogeny and correspond to the major geographical regions in the study: Africa, Brazil (2 clades), the UK, and the US (figure 2). The geographical clustering is conserved using both core gene (figure 1), core kmer (supplemental figure 1), and ancestral state reconstruction through BEAST2 (14)(figure 2). Figure 1 is rooted to a reference *S.* Enteritidis (AM933172). *S.* Enteritidis was chosen as an outgroup of *S.* Dublin based upon the serotype phylogeny provided by SISTR (https://lfz.corefacility.ca/sistr-app/). Some host preference clustering is evident in the African and Brazilian clades as the genomes are predominately of human origin (outer colored ring, figure 1). However, it is likely that such clustering is a consequence of sampling bias in the publicly available genome datasets (table 1) and is a by-product of geographical clustering. Examination of the US clade, which is roughly split between bovine, food, and human genomes reveals no monophyletic groups regarding host source. The same scenario is seen in the UK clade which is more balanced in terms of isolation source. Thus, the core genome sequence of *S.* Dublin is highly influenced by the region the genomes originates, and not the isolation host.

To further explore the population structure, ancestral state reconstruction was conducted and plotted in a simplified phylogeny (figure 2). The five major clades observed using core gene maximum-likelihood methods are conserved and collapsed for clarity. Consistent with figure one, the tree is rooted to *S.* Enteritidis AM933172. The first divergence event in the phylogeny is the emergence of a single outgroup (SRR1122707)

that diverges from all other *S.* Dublin genomes. The genome was isolated from a bovine source in 1982 in France. However, due to the sparsity of provenance, and the focus on population genomics, no conclusions can be made as to why the genome is divergent from the dataset. It does pose an interesting possibility that a unique clade of *S.* Dublin is grossly undersampled in the current database. The next divergence event, or emergence, was the African clade, followed by the minor Brazilian clade. The final and largest divergence event yield two major clades, the US clade, and a mixed clade populated with the major Brazilian and UK clade. The two Brazilian clades emerged at different evolutionary time points revealing two distinct lineages that were present in Brazil. Additionally, the major Brazilian clade appears to have UK origins. The UK and Brazilian clade share a common ancestor with a UK genome SRR6191689. The monophyletic group, consisting of the UK, major Brazilian clade, and SRR6191689, diverged from a common ancestor shared by four UK and one US genomes. Thus, the major Brazilian clade was probably introduced to the country from the UK. However, such an explanation cannot be extended to the minor Brazilian clade, whose origin is unclear. The US clade shares a common ancestor with a single Brazilian genome, SRR6701626 and the combined clade shares a common ancestor with six UK genomes. At this time point, a definitive statement to the origin of the US genomes cannot be made.

Ancillary genome composition is geography dependent.

Core genome composition is more indicative of geographical location than host source. The next possibility we decided to investigate was the influence of geography on the ancillary or accessory genome (defined here as 5% < gene prevalence < 99%). Genes with a prevalence less than five percent were excluded to minimize the confounding

effects of improper open reading frame identification and sequencing error. Ancillary gene catalogues, consistent with the core genome composition, are influenced by geography (figure 3). Logistic PCA, an extension to classical PCA used to reduce dimensionality in binary matrices, was used to plot genomes in a two-dimensional space respective of their ancillary genome content. Two large clusters of genomes are clearly represented in figure 3: the majority of US genomes cluster to the left of the plot whereas most of the global and a smaller number of US genomes clustering to the right. Figure 3 illustrates that US genomes harbor an ancillary genomic catalogue that readily distinguishes the genomes from isolates not originating from the US. Remarkably, sub-clustering within the right cluster separates genomes into the five clades witnessed in the core genome phylogeny (figure 1,2). Ancillary genome composition, as core genome structure, is a geographical characteristic of *S.* Dublin genome.

Further work was carried out to determine what genomic elements were responsible for region-specific differentiation of genomes. The pangenome of *S.* Dublin was constructed and is presented in figure 4A. Stated earlier, the core genome of *S.* Dublin is 4,098 genes. Genes with a prevalence of $5\% < x < 99\%$ (shell genes) numbered 833 genes. Lastly, genes with a prevalence of less than 5% (cloud genes) numbered 5,533 for a total pangenome size of 10,464 non-redundant genes. Interestingly, the number of core genes shared by 99% of genomes was nearly five times as great as genes shared between 15% to 99% of genomes. Such a disparity highlights high conservation of core genes. Put another way, much of the gene increase in the pangenome is due to the addition of unique coding sequences into a small number of genomes. However, a major exception to this statement can be seen in figure 4A. The shell pangenome is plotted as a

binary matrix against the phylogeny of the serotype. A block of nearly 100 genes is clearly seen and is found only in US genomes. Said block of genes is responsible for the clustering pattern in figure 3 where the US genomes cluster away from global genomes. Examination of the aforementioned block showed genes pertaining to antibiotic resistance, toxin-antitoxin systems, and a large number of hypothetical or uncharacterized proteins. US genomes yield larger assemblies (figure 3 B,C) and more predicted open reading frames (figure 3 D,E) due to this gene block.

A possible explanation for the large increase in assembly size and number of open reading frames in US genomes is the acquisition of mobile genetic elements (prophages, plasmids, etc.,). We arrived at such a hypothesis by the presence of toxin-antitoxin systems in the unique US gene block. To investigate the said possibility, prophage regions and plasmid replicons were identified in genomes using the web service PHASTER and the PlasmidFinder database (please see methods). Prophage insertions are not responsible for the increased gene number in US genes and are not a major diversifying agent in the serotype (figure 5A, right panel). The full PHASTER phage details are provided in supplemental data set 2. Three major phages were identified in the *S. Dublin* queried: Gifsy 2, sal3, and RE 2010. The prevalence of the three major phages is greater than 90% for all regions (figure 5B) and no US specific or any region-specific pattern is present. However, plasmid replicons did yield a region-specific pattern. IncA/C2 replicon was identified only in US genomes. A representative contig containing IncA/C2 replication site was extracted from the genome assembly of SRR5000235. The sequence yielded a 99% sequence identity (BLAST+) to an IncA/C antibiotic resistance plasmid isolated from Salmonella Newport (CP009564.1). The other major replicon

identified in *S.* Dublin was IncX1. Extracting the contig sequences with replicon and BLAST search yielded 99% sequence identity to the *S.* Dublin virulence plasmid (CP032450.1). The replicon is highly conserved among the genomes and is a common characteristic of *S.* Dublin. IncX1 was identified in 865 (98%) of *S.* Dublin genomes followed by IncFII(S) identified in 826 (94%) genomes and IncA/C2 identified in 476 (54%) genomes. Thus plasmids, not prophages, diversify *S.* Dublin. The large block of US genes, and resulting gene count and assembly length increase are due to the presence of an IncA/C2 resistance plasmid found only in US genomes.

Antimicrobial resistance is a US phenomenon.

Antibiotic resistance is a characteristic of *S.* Dublin in the United States, but not a characteristic of the serotype. Many of the US genomes contain multiple predicted AMR homologues as shown in figure 6A. The matrix reveals that AMR homologues are largely absent from genomes that were not isolated in the US. The bimodal distribution of AMR homologues is clearly shown in figure 6B: The median AMR homologue per genome in the US was 5, with all other regions yielding a median value of zero. The most abundant classes of antibiotics that the serotype is resistant to (figure 6C) are : aminoglycosides, beta-lactams, phenicols, sulfonamides, and tetracyclines. Importantly, no resistance homologues to quinolones or fluoroquinolones were detected. Full details of the AMR identification can be found in supplemental data set 3. One possibility for the increase of AMR genes in the US genomes may be time; many of the US genomes were recently isolated. However, date of isolation is not a significant factor and does not explain the increase of AMR genes (figure 6D). Many international isolates collected at similar time points yield no AMR homologues. Thus, as seen with the core genome

sequence variation and ancillary gene content, AMR homologues are a geographical phenomenon.

S. Dublin and S. Enteritidis harbor unique pangenomes

The final investigation was conducted to define which genomic features if any, define *S.* Dublin as a serotype. To accomplish this, 160 genomes of *S.* Enteritidis, the closest known serological neighbor of *S.* Dublin, were download from the sequence read archive hosted by NCBI. Briefly, the Pathogen Detection metadata, previously downloaded, was queried for *S.* Enteritidis. From the resultant list, 160 were randomly sub-sampled to include genomes from Europe, North America, Asia, and Africa. Genome assemblies were validated for assembly quality and serotyping prediction consistent with the core *S.* Dublin dataset. Core gene phylogeny was estimated and readily separated the serotypes (supplemental figure 3) into serotype specific clades. Furthermore, distinct blocks of genes originating from the two serotypes were observed in combined pangenome (figure 7A). *S.* Dublin and *S.* Enteritidis genomes shared 3,760 genes. Shell genes, prevalence $15\% < x < 99\%$, numbered 1,057. The total pangenome for the two serotypes was composed of 13,835 genes. In addition to core genome variation, ancillary gene content also separates the two serotypes (figure 7B). Thus, core and ancillary genomic features between *S.* Dublin and *S.* Enteritidis are distinct. Identification of core *S.* Dublin and *S.* Enteritidis genes was defined simply as: abs(Dublin_prevalence – Enteritidis_prevalence) > 0.99. Using said criteria, as well as the software Scoary (15), 82 *S.* Dublin specific genes were identified. Additionally, 30 *S.* Enteritidis specific genes were identified. The full gene list with manual annotations, significance values, and BLASTP accession numbers are provided in supplemental data set 4. *S.* Dublin and *S.* Enteritidis specific

genes are illustrated in a binary matrix grouped by functional category shown in figure 8. The largest functional differences between the serotypes are genes that code for phage products, transporters, metabolic pathways, and hypothetical proteins. *S.* Dublin specific metabolic genes include sugar dehydrogenases (glucose, soluble aldose sugars, and glucarate) and propionate catabolism regulatory proteins. In addition to specific sugar catabolic genes, multiple PTS and ABC transporters were identified as *S.* Dublin specific genes. Accordingly, *S.* Dublin specific pathways code for the transport and catabolism of carbohydrates. *S.* Dublin contains two virulence genes not found in *S.* Enteritidis, type VI secretion protein VgrG and fimbrial protein subunit FimI.

DISCUSSION

In the work presented, we establish the global population structure of *S.* Dublin. Geography exerts a strong influence on the core (figure 1, 2) and the ancillary genome (figure 3,4). Region-specific clades dominate the global population structure of *S.* Dublin. Strains of *S.* Dublin in the UK are genomically distinct from US strains (and distinct from Brazilian and African, etc.). Such differences and ancestral state reconstruction suggest a vicariant model of evolution. The major Brazilian clade was most likely introduced from the UK. The clade shares a common lineage with 5 UK genomes as well as the UK clade. It has been suggested that the UK acts a source of *S.* Dublin dissemination to distant populations such as South Africa and Australia (16). Once introduced into the new geographical region, the strains began to diversify from the parental UK strains in both core and ancillary genome composition. Ancillary gene catalogues are distinct enough between regions to allow clustering based solely upon presence-absence matrices. Geographically distinct strains have been identified in *Salmonella enterica* Typhimurium

4,[5],12:i:- where strains isolated from similar areas form monophyletic groups (17). Similar phylogeographical separation has also been observed in *S.* Dublin as well. Strains isolated from New York and Washington states cluster into distinct clades (9). We did not consider within country clustering due to the paucity of certain samples metadata (lack of specific region details). Strong phylogeographical clustering may be explained by the pathogenesis and host preference of *S.* Dublin. Cattle is the primary host of *S.* Dublin and the establishment of a carrier state has been implemented in the maintenance of herd infections (18, 19). Indeed, it has been shown that the geographical clustering of *S.* Dublin infected herds is strongly associated with cattle movement patterns in Norway (20) compared to *S. enterica* Typhimurium. The authors suggest that *S. enterica* Typhimurium can utilize multiple hosts for dispersion, whereas *S.* Dublin is largely relegated to herd movement. Such dependence on host movement, and host movement dependence on agricultural practices, could explain why *S.* Dublin is independently evolving around the globe: exposure to susceptible populations is limited.

AMR homologues are a US phenomenon associated with the IncA/C2 plasmid replicon. IncA/C conjugative plasmids are typically isolated from Enterobacteriaceae. It has been suggested that the plasmid was first acquired from an environmental source and gained antibiotic resistance homologues and systems in response to agricultural selective pressures (21). Indeed, it has been shown *in vitro* and *in vivo* calf dairy models that IncA/C plasmid carriage exerts a measurable negative fitness cost upon the host bacterium and without selective pressure, the host will cure themselves of the plasmid (22). *S.* Dublin is primarily associated with dairy cattle and can establish an asymptomatic carrier state. It is reasonable to assert that antibiotics given to a carrier cow

to treat another condition would satisfy the selective pressure required to ensure IncA/C is retained. Mastitis is the primary condition for which dairy cows receive antibiotic treatment (23) and many dairy cows receive antimicrobial treatment following lactation to prevent mastitis (24). What is more alarming however, is the discovery of a large (172, 265 bp) hybrid plasmid combing the *S.* Dublin virulence plasmid to the IncA/C2 plasmid (25). The authors note that the new hybrid plasmid pN13-01125 yields resistance homologues to at least six classes of antimicrobial agents and a low conjugation frequency. However, the plasmid is reliably inherited to daughter cells. The inclusion of the IncA/C2 plasmid into the main virulence plasmid of *S.* Dublin will increase the stability of the genes and could represent a scenario where the main virulent factors of systemic infection are intimately tied with the AMR gene of US isolates. The study identifying the hybrid plasmid did so through the aid of long-read-length sequencing on the Pacific Biosciences RSII system. Our study relied upon short-read sequencing and assembly. Thus, it is difficult to ascertain the presence or absence of the hybrid plasmid as the sequence could be fragmented into multiple contigs. The future evolution of the IncA/C2 plasmids and hybrid plasmids in *S.* Dublin will need to be studied further with the need for IncA/C2 specific PCR assay development.

Previous studies have reported genomic variances between *S.* Dublin and *S. Enteritidis* (26-28). Indeed, using DNA microarrays it was determined that *S.* Dublin contains 87 specific genes and *S. Enteritidis* contains 33 serotype specific genes (26). Said work was conducted comparing 4 *S.* Dublin against a set of 29 *S. Enteritidis*. Even with a reduced number of genomes, the results strongly agree with our findings in that we identified 82 and 30 *S.* Dublin and *S. Enteritidis* specific genes respectively from a set of

880 *S.* Dublin against 160 *S. Enteritidis*. One of the prominent *S.* Dublin specific genes identified in this work and others is the Gifsy-2 prophage. It has been shown that deletion of the Gifsy-2 phage in *S. enterica* Typhimurium significantly decreases the organisms ability to establish systemic infections in mice (29) and has been recently identified as part of the *S.* Dublin invasome (30). In addition to the Gifsy-2 phage, two additional virulence factors were identified as *S.* Dublin specific: a type VI secretion protein VgrG and the type I fimbral subunit FimI. VgrG, encoded by Salmonella pathogenicity island 19 (SPI-19) aids in macrophage survival in the host-adapted serotype *S. enterica* Gallinarum (31). Intact SPI-19 has been isolated in *S. Enteritidis*, however "classical" isolates of the serotype, those which commonly infect humans and animals, contain a degraded version of SPI-19 (11). That some *S. Enteritidis* encode the full SPI-19 suggests that *S.* Dublin has not gained said virulence gene, rather maintained them after the divergence from *S. Enteritidis*. Metabolic genes and transporters were additionally found to be *S.* Dublin specific. Many of the *S.* Dublin specific metabolic genes encoded the uptake and catabolism of carbohydrates. Such metabolic pathways may be advantageous for survival in the rumen environment. Competent *S.* Dublin cells have been cultured from the rumen fluid of slaughter cows (32) showing the bacterium is capable of surviving the rumen. However, due to the complexity of *S.* Dublin's virulence, coupled with the high number of hypothetical proteins, wet lab work will be needed to define the importance of many of the *S.* Dublin specific genes.

Figures and Tables

TABLE 1 Metadata characteristics of 880 *S.* Dublin comprising the dataset. The table is grouped into the four major geographical regions the genomes originate.

Region	Source	Num. Genomes
Africa	Bovine	7
	Food	2
	Human	26
Brazil	Bovine	24
	Human	89
	Other	4
United Kingdom	Food	23
	Human	99
	Other	56
United States	Bovine	231
	Food	189
	Human	118
	Other	11

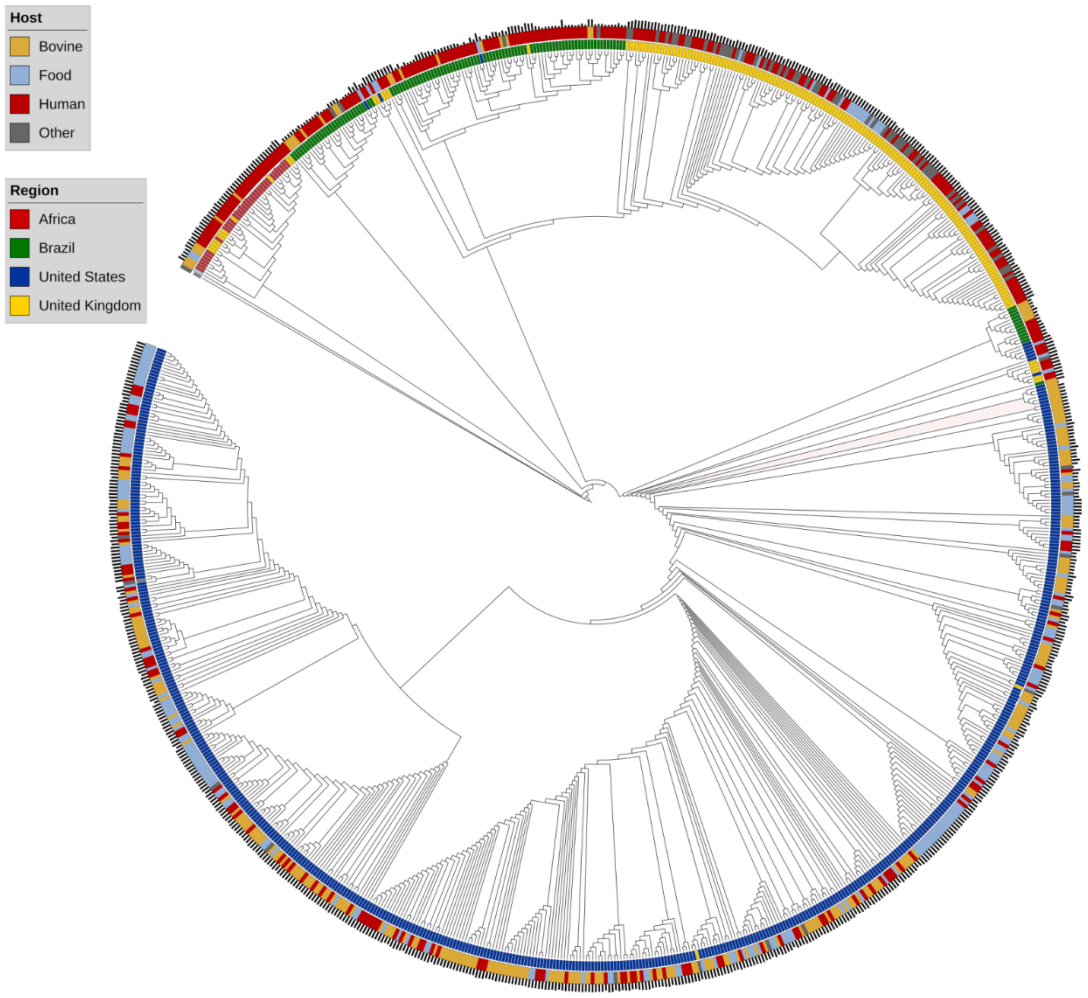


FIG 1 Global population structure of *S.* Dublin illustrated with a maximum-likelihood cladogram, GTR Gamma model, of 880 *S.* Dublin genomes rooted to *S.* Enteritidis (AM933172). The tree is inferred from an alignment of 4,098 core genes defined by Roary. Leaves are colored respective to the region of isolation. The outer colored ring denotes isolation source. The outermost barplots illustrate data of isolation and are scaled such that the higher the bar, the more recent the isolate was cultured and zeroed to a date of 1980. Genomes cluster into distinct clades associated with the area of isolation.

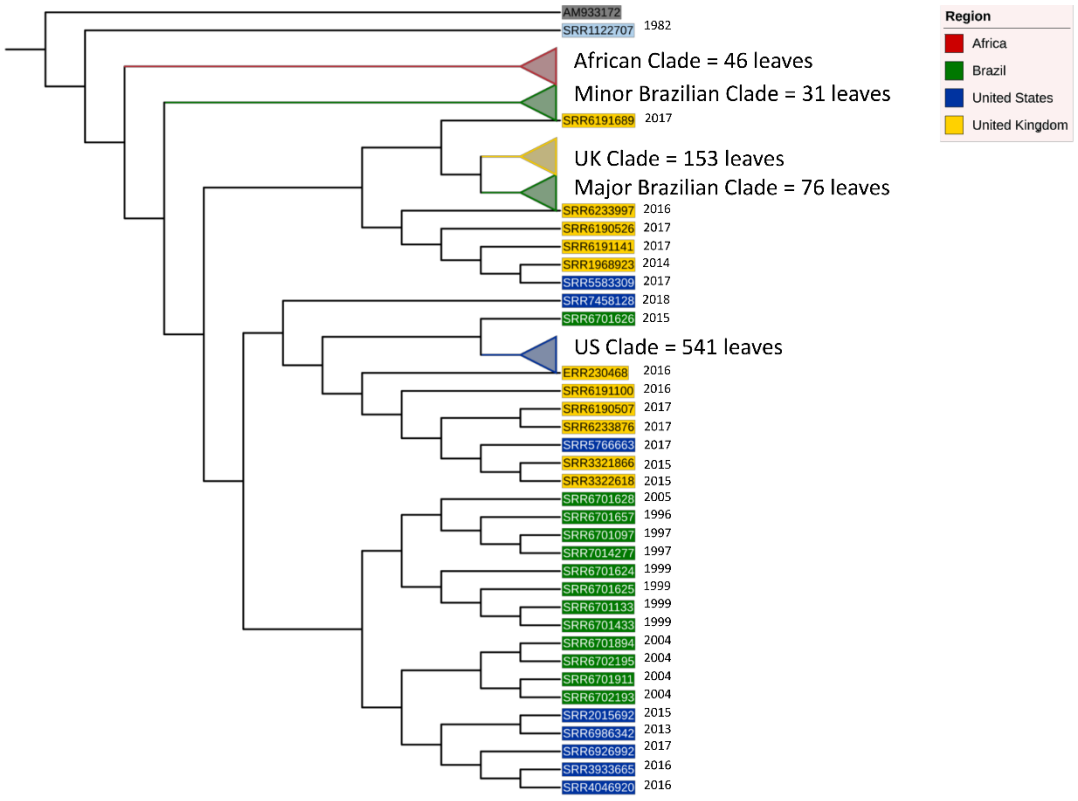


FIG 2 Reduced phylogeny of *S.* Dublin. Ancestral state reconstruction using BEAST2 on core genes conserved major geographical clades. Clades are collapsed to aid in visualization of high-order tree architecture. The phylogeny is rooted to *S.* Enteritidis AM933172. Clades and leaf labels are colored respective to the region of isolation. Isolation date is listed to the right of leaves not collapsed into clades.

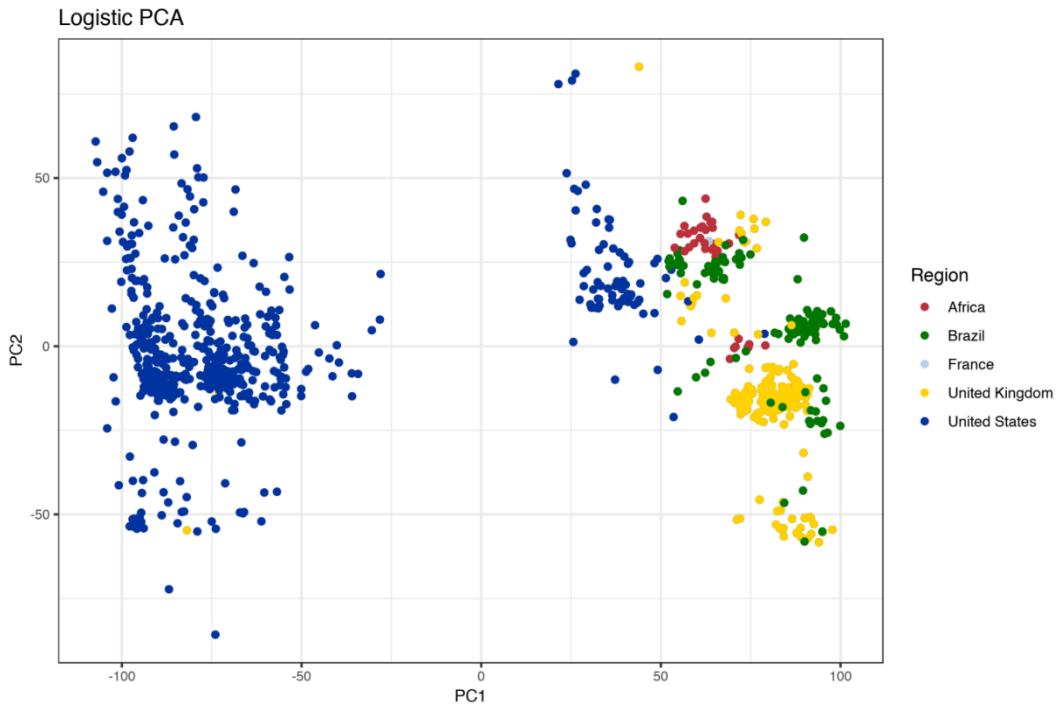


FIG 3 Ancillary genes cluster *S.* Dublin geographically. Logistic PCA was run on a binary matrix of ancillary genes, prevalence less than 99% and greater than 5%. Region-specific clusters appear and correspond to the major geographical clades. Additionally, most of the US genomes cluster away from global genomes.

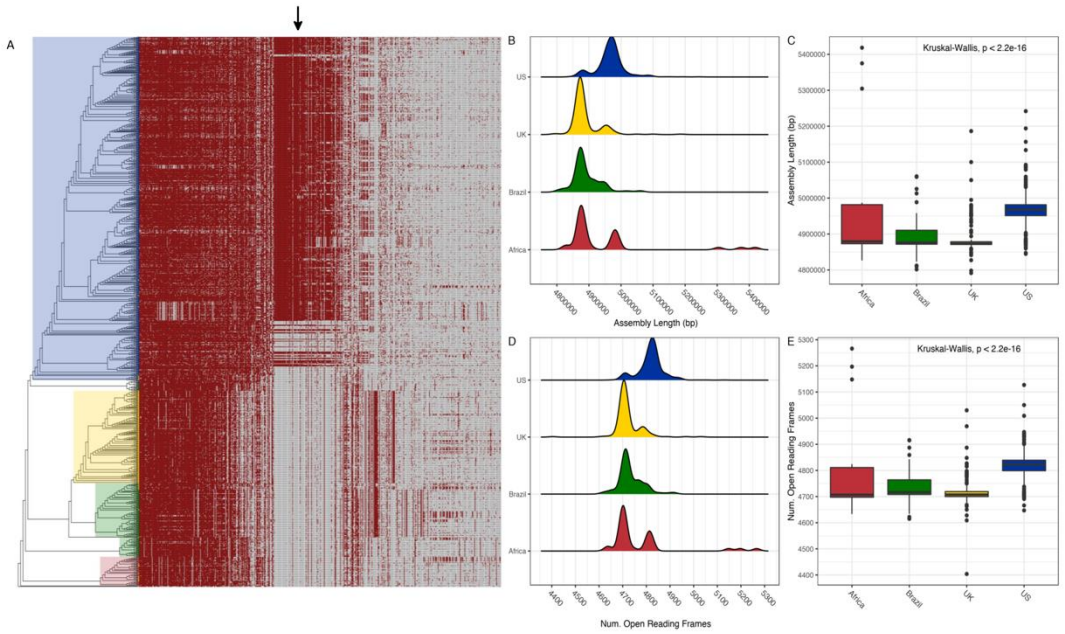
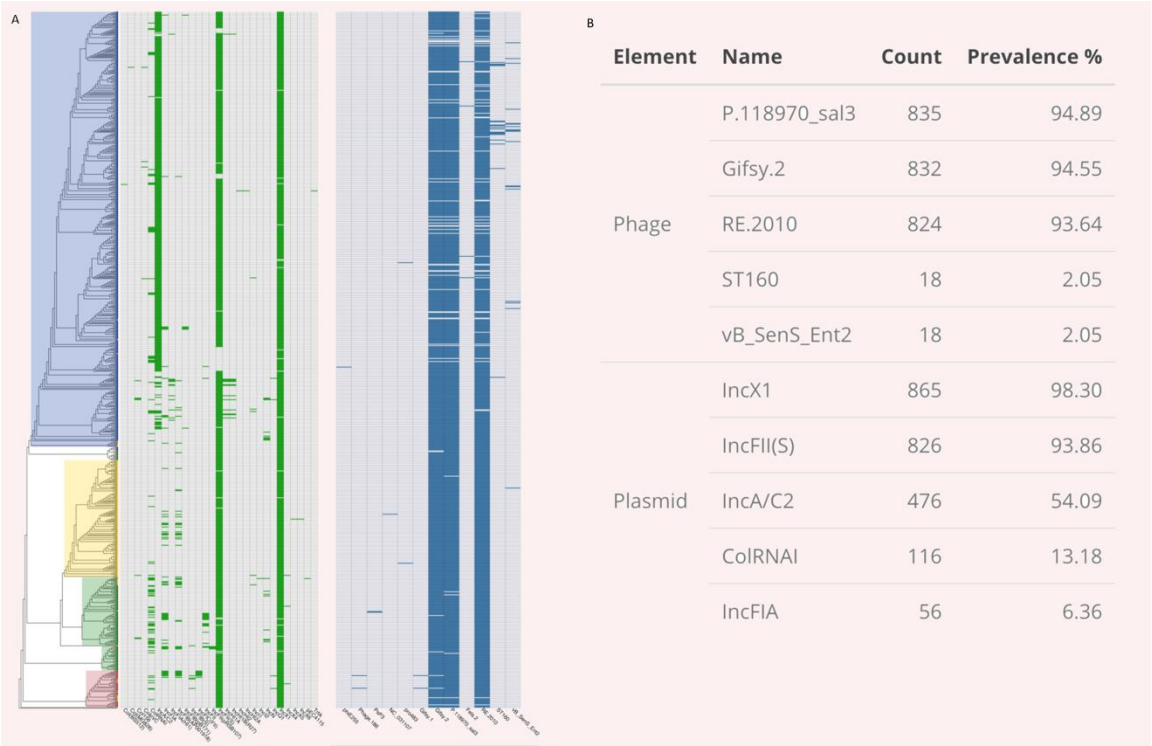


FIG 4 Pangenome of *S.* Dublin. (A) Presence-absence matrix of the pangenome plotted against the phylogeny of *S.* Dublin. Note that core genes and genes with a prevalence of less than 5% were removed to enhance clarity. The arrow denotes a block of genes linked with the IncA/C2 plasmid replicon. (B)(C) Density and bar plots of assembly length plotted according to the region of isolation. US genomes are approximately 100kb longer than genomes from other regions. (D)(E) US genomes contain approximately 100 more open reading frames than genomes from other regions. Density plots are colored respective to region of isolation consistent with figures 2 and 3.



Element	Name	Count	Prevalence %
Phage	P.118970_sal3	835	94.89
	Gifsy.2	832	94.55
	RE.2010	824	93.64
	ST160	18	2.05
	vB_SenS_Ent2	18	2.05
Plasmid	IncX1	865	98.30
	IncFII(S)	826	93.86
	IncA/C2	476	54.09
	ColRNAI	116	13.18
	IncFIA	56	6.36

FIG 5 Mobile genetic elements of *S.* Dublin. (A) Multi-panel matrices illustrating the presence-absence of plasmid replicons (left) and prophages (right) identified in *S.* Dublin aligned to the phylogeny. Phage content is conserved among the serotype. Plasmid replicons show more varied distribution. InxC1, corresponding to the *S.* Dublin virulence plasmid is highly conserved. IncA/C2, homologous to a *S.* Newport resistance plasmid, is found only in US genomes. (B) Table describing the top 5 most abundant plasmid replicons and prophages identified in the dataset.

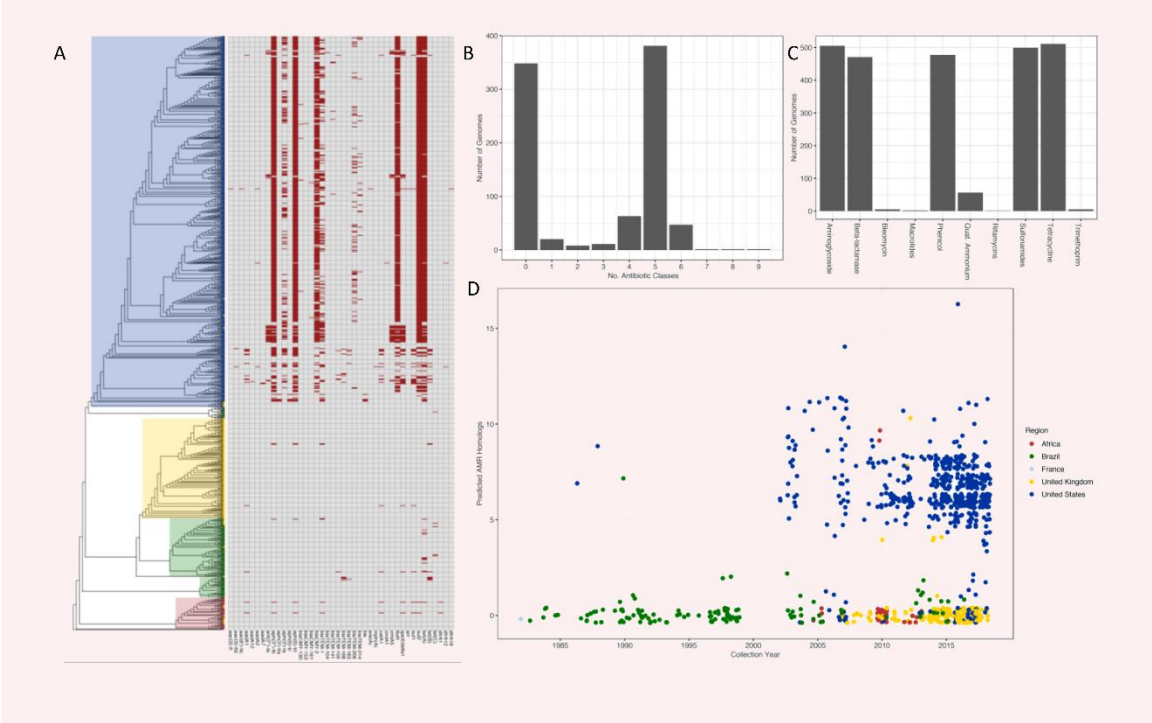


FIG 6 Antimicrobial resistance (AMR) patterns of *S.* Dublin are geography dependent. (A) AMR homologue presence-absence for individual genomes plotted against the phylogeny of *S.* Dublin. Genomes in the US contain more predicted resistance homologues than global genomes. (B) *S.* Dublin shows a bimodal distribution of antibiotic resistance where genomes are either contain no predicted homologues or homologues conferring resistance to five classes of antibiotics. (C) The five most abundant classes of AMR homologues found in the genomes: aminoglycosides, beta-lactams, phenicols, sulfonamides, and tetracycline. (D) Number of AMR homologues per genome plotted in relation to the year of collection. Between the period of 2000 – 2018, US genomes contain more AMR homologues than genomes from other regions. Dots represent individual genomes colored respective to the area of isolation.

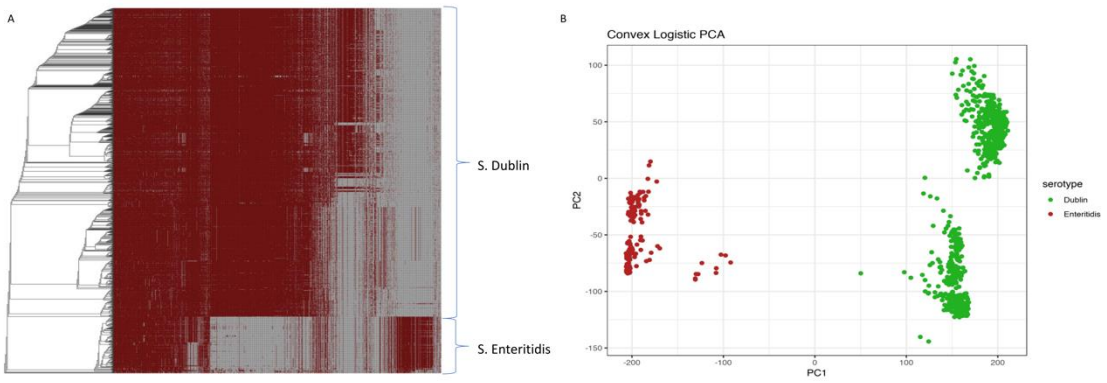


FIG 7 *S.* Dublin and *S.* Enteritidis differ in pangenome composition. (A) Gene presence-absence matrix plotted against the phylogeny of *S.* Dublin and *S.* Enteritidis. Unique sets of genes can be observed in the *S.* Dublin and *S.* Enteritidis clades of the matrix. (B) Ancillary gene content differentiates *S.* Dublin from *S.* Enteritidis. Each dot represents a genome and is colored respective to serotype. Three large clusters are shown depicting a split between *S.* Dublin and *S.* Enteritidis. Logistic PCA was conducted on genes with a prevalence of less than 99% and greater than 5%.

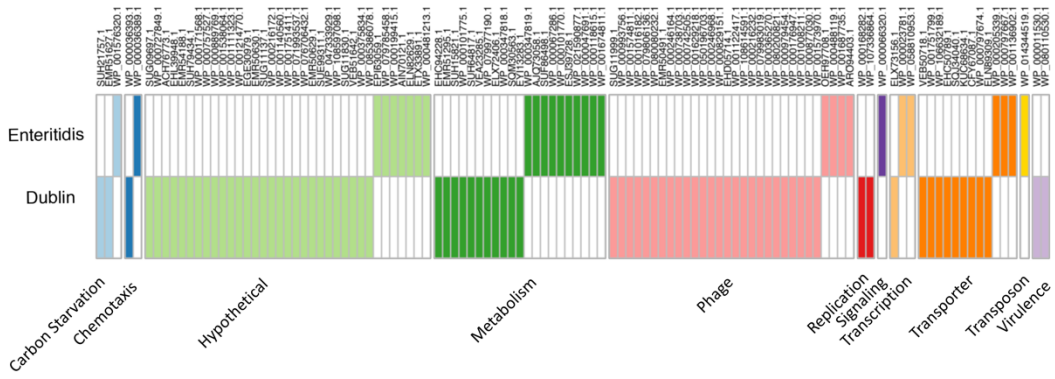


FIG 8 *S.* Dublin and *S.* Enteritidis serotype specific genes. Genes are grouped by functional annotation and color denotes presence, white denotes gene absence. 82 genes were identified as *S.* Dublin specific and 30 genes were identified as S. Enteritidis specific. Column names correspond to the closet homologue determined by manual curation.

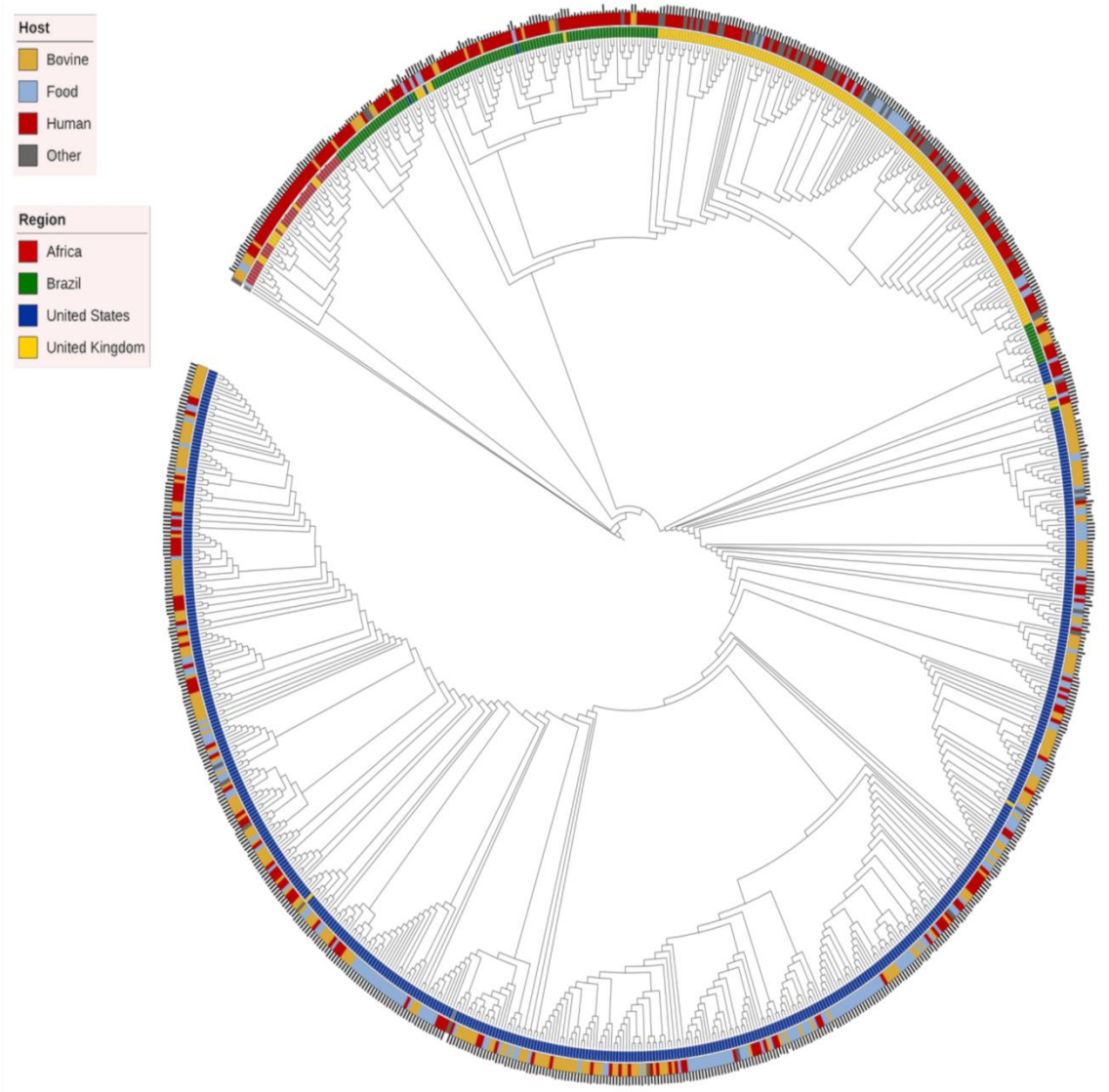


FIG S1 Population structure of *S.* Dublin identified using core kmer content with the software KSNP3. Leaves are colored respective to the region of isolation. The outer ring is colored respective to isolation source and the outer bars are scaled to date of isolation. The higher the bar the more recent the isolate was cultured. Strong geographical clustering is observed with clades corresponding to regions of isolation.

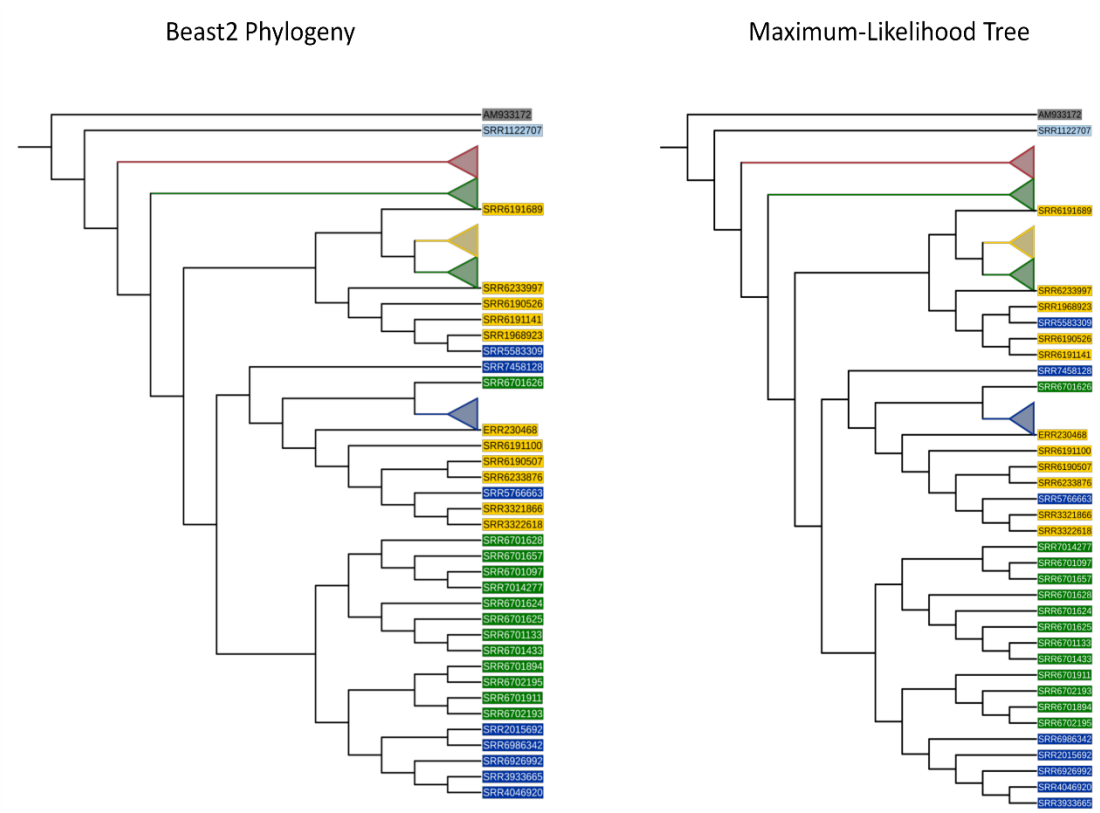


FIG S2 Comparison of the ancestral state phylogeny with the maximum-likelihood phylogeny. Major clades are collapsed for comparison purposes. The five major clades are conserved in both phylogenomic methods.

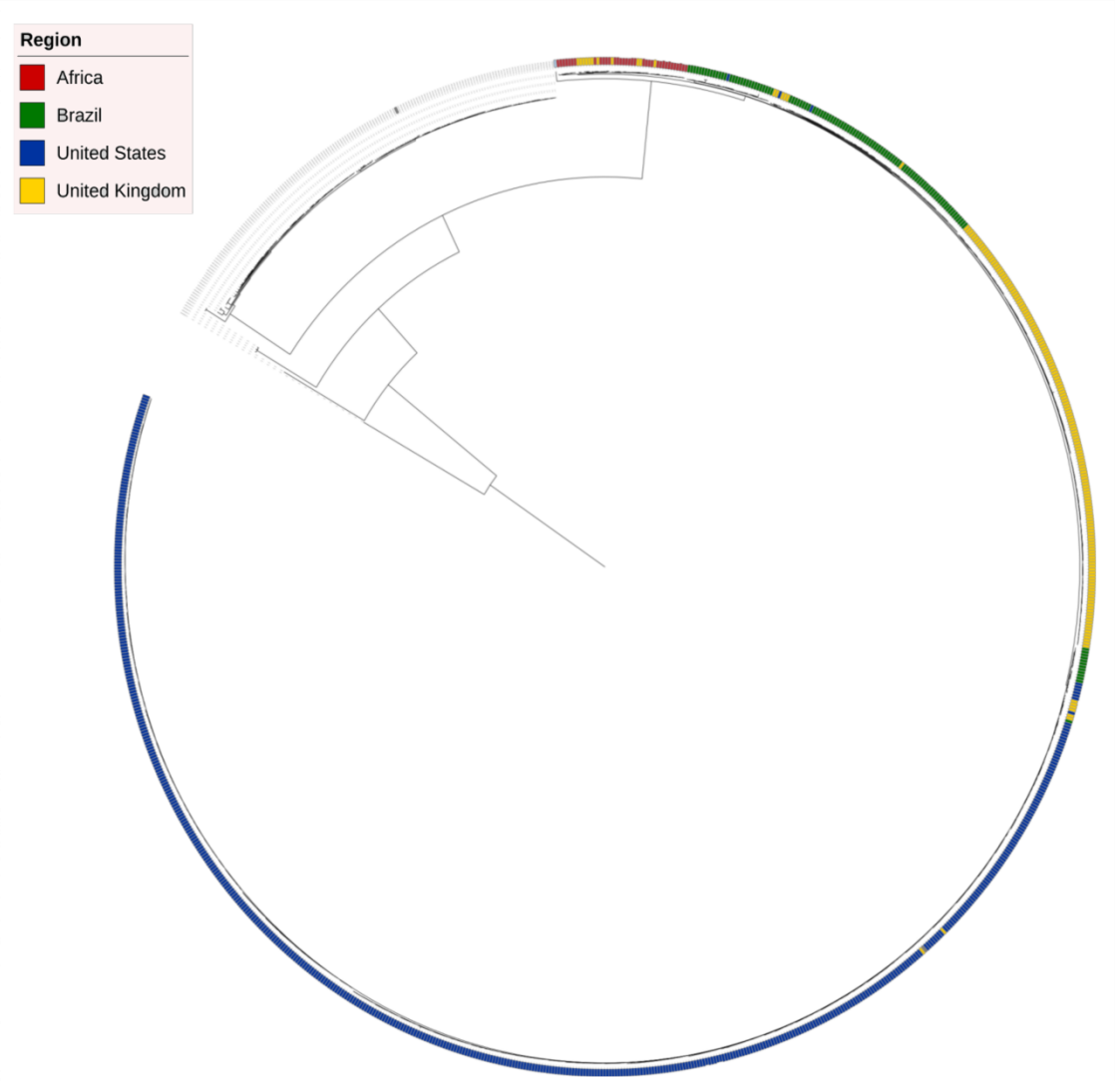


FIG S3 Maximum-likelihood tree showing the phylogeny of 880 *S.* Dublin and 161 S. Enteritidis. *S.* Dublin leaves are colored respective to the region of isolation and one *S.* Enteritidis (AM933172) that was used to root the *S.* Dublin phylogeny. *S.* Dublin forms a single clade away from *S.* Enteritidis. Geographical clades are seen in the *S.* Dublin clade.

REFERENCES

1. Wallis TS, Paulin SM, Plested JS, Watson PR, Jones PW. 1995. The Salmonella dublin virulence plasmid mediates systemic but not enteric phases of salmonellosis in cattle. Infection and Immunity 63:2755.
2. Nielsen LR. 2013. Review of pathogenesis and diagnostic methods of immediate relevance for epidemiology and control of Salmonella Dublin in cattle. Veterinary Microbiology 162:1-9.
3. Pullinger GD, Paulin SM, Charleston B, Watson PR, Bowen AJ, Dziva F, Morgan E, Villarreal-Ramos B, Wallis TS, Stevens MP. 2007. Systemic Translocation of Salmonella enterica Serovar Dublin in Cattle Occurs Predominantly via Efferent Lymphatics in a Cell-Free Niche and Requires Type III Secretion System 1 (T3SS-1) but Not T3SS-2. Infection and Immunity 75:5191-5199.
4. Libby SJ, Adams LG, Ficht TA, Allen C, Whitford HA, Buchmeier NA, Bossie S, Guiney DG. 1997. The spv genes on the Salmonella dublin virulence plasmid are required for severe enteritis and systemic infection in the natural host. Infection and Immunity 65:1786.
5. Nielsen LR, Schukken YH, Grohn YT, Ersboll AK. 2004. Salmonella Dublin infection in dairy cattle: risk factors for becoming a carrier. Prev Vet Med 65:47-62.
6. Maguire H, Cowden J, Jacob M, Rowe B, Roberts D, Bruce J, Mitchell E. 1992. An outbreak of Salmonella dublin infection in England and Wales associated with a soft unpasteurized cows' milk cheese. Epidemiology and Infection 109:389-396.

7. Werner SB, Humphrey GL, Kamei I. 1979. Association between raw milk and human Salmonella dublin infection. British Medical Journal 2:238.

8. Small RG, Sharp JCM. 1979. A milk-borne outbreak due to Salmonella dublin. Journal of Hygiene 82:95-100.

9. Carroll LM, Wiedmann M, den Bakker H, Siler J, Warchocki S, Kent D, Lyalina S, Davis M, Sischo W, Besser T, Warnick LD, Pereira RV. 2017. Whole-Genome Sequencing of Drug-Resistant Salmonella enterica Isolates from Dairy Cattle and Humans in New York and Washington States Reveals Source and Geographic Associations. 83:e00140-17.

10. Matthews TD, Schmieder R, Silva GGZ, Busch J, Cassman N, Dutilh BE, Green D, Matlock B, Heffernan B, Olsen GJ, Farris Hanna L, Schifferli DM, Maloy S, Dinsdale EA, Edwards RA. 2015. Genomic Comparison of the Closely-Related Salmonella enterica Serovars Enteritidis, Dublin and Gallinarum. PLOS ONE 10:e0126883.

11. Langridge GC, Fookes M, Connor TR, Feltwell T, Feasey N, Parsons BN, Seth-Smith HMB, Barquist L, Stedman A, Humphrey T, Wigley P, Peters SE, Maskell DJ, Corander J, Chabalgoity JA, Barrow P, Parkhill J, Dougan G, Thomson NR. 2015. Patterns of genome evolution that have accompanied host adaptation inSalmonella. Proceedings of the National Academy of Sciences 112:863-868.

12. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VPJ, Nash JHE, Taboada EN. 2016. The Salmonella In Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft Salmonella Genome Assemblies. PLOS ONE 11:e0147101.

13. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17:132.

14. Bouckaert R, Heled J, Kuhnert D, Vaughan T, Wu CH, Xie D, Suchard MA, Rambaut A, Drummond AJ. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput Biol 10:e1003537.

15. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. 2016. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biology 17.

16. Felsenfeld OY, V. M. 1947. The Geography of Salmonella. American Journal of Digestive Diseases 14:47-52.

17. Palma F, Manfreda G, Silva M, Parisi A, Barker DOR, Taboada EN, Pasquali F, Rossi M. 2018. Genome-wide identification of geographical segregated genetic markers in Salmonella enterica serovar Typhimurium variant 4,[5],12:i. Scientific Reports 8.

18. Henderson K, Mason C. 2017. Diagnosis and control of Salmonella Dublin in dairy herds. In Practice 39:158-168.

19. McGuirk SMP, Simon. Salmonellosis in Cattle : A Review, p. *In* (ed),

20. Lewerin SS, Skog L, Frössling J, Wahlström H. 2011. Geographical distribution of salmonella infected pig, cattle and sheep herds in Sweden 1993-2010. Acta Veterinaria Scandinavica 53:51.

21. Johnson TJ, Lang KS. 2012. IncA/C plasmids. Mobile Genetic Elements 2:55-58.

22. Subbiah M, Top EM, Shah DH, Call DR. 2011. Selection Pressure Required for Long-Term Persistence ofblaCMY-2-Positive IncA/C Plasmids. Applied and Environmental Microbiology 77:4486-4493.

23. Saini V, McClure JT, Léger D, Dufour S, Sheldon AG, Scholl DT, Barkema HW. 2012. Antimicrobial use on Canadian dairy farms. Journal of Dairy Science 95:1209-1221.

24. Landers TF, Cohen B, Wittum TE, Larson EL. 2012. A Review of Antibiotic Use in Food Animals: Perspective, Policy, and Potential. Public Health Reports 127:4-22.

25. Mangat CS, Bekal S, Irwin RJ, Mulvey MR. 2017. A Novel Hybrid Plasmid Carrying Multiple Antimicrobial Resistance and Virulence Genes in Salmonella enterica Serovar Dublin. Antimicrobial Agents and Chemotherapy 61:e02601-16.

26. Betancor L, Yim L, Martínez A, Fookes M, Sasias S, Schelotto F, Thomson N, Maskell D, Chabalgoity JA. 2012. Genomic Comparison of the Closely Related Salmonella enterica Serovars Enteritidis and Dublin. The open microbiology journal 6:5-13.

27. Porwollik S, Santiviago CA, Cheng P, Florea L, Jackson S, McClelland M. 2005. Differences in gene content between Salmonella enterica serovar Enteritidis isolates and comparison to closely related serovars Gallinarum and Dublin. J Bacteriol 187:6545-55.

28. Mohammed M, Cormican M. 2016. Whole genome sequencing provides insights into the genetic determinants of invasiveness in Salmonella Dublin. Epidemiol Infect 144:2430-9.

29. Figueroa-Bossi N, Bossi L. 1999. Inducible prophages contribute to Salmonella virulence in mice. Molecular Microbiology 33:167-176.

30. Mohammed M, Le Hello S, Leekitcharoenphon P, Hendriksen R. 2017. The invasome of Salmonella Dublin as revealed by whole genome sequencing. BMC Infect Dis 17:544.

31. Blondel CJ, Jiménez JC, Leiva LE, Álvarez SA, Pinto BI, Contreras F, Pezoa D, Santiviago CA, Contreras I. 2013. The Type VI Secretion System Encoded in Salmonella Pathogenicity Island 19 Is Required for Salmonella enterica Serotype Gallinarum Survival within Infected Macrophages. Infection and Immunity 81:1207-1220.

32. McEvoy JM, Doherty AM, Sheridan JJ, Blair IS, McDowell DA. 2003. The prevalence of Salmonella spp. in bovine faecal, rumen and carcass samples at a commercial abattoir. Journal of Applied Microbiology 94:693-700.

33. Torsten Seemann JK, Simon Gladman, Ander Goncalves da Silva. Shovill : Assemble bacterial isolate genomes from Illumina paired-end reads, GitHub, GitHub. https://github.com/tseemann/shovill.

34. Song L, Florea L, Langmead BJGB. 2014. Lighter: fast and memory-efficient sequencing error correction without counting. 15:509.

35. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455-77.

36. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One 9:e112963.

37. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068-9.

38. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691-3.

39. Loytynoja A. 2014. Phylogeny-aware alignment with PRANK. Methods Mol Biol 1079:155-70.

40. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. 2:-.

41. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242-5.

42. Gardner SN, Slezak T, Hall BG. 2015. kSNP3.0: SNP detection and phylogenetic analysis of genomes without genome alignment or reference genome. Bioinformatics 31:2877-8.

43. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312-1313.

44. Seemann T. 2018. ABRicate, v0.8.0. GitHub. https://github.com/tseemann/abricate.

45. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. 2014. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrobial agents and chemotherapy 58:3895-3903.

46. Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis, 2 ed. Springer Publishing Company, Incorporated

47. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y, McInerny G. 2017. ggtree: anrpackage for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 8:28-36.

48. Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics 32:2847-9.

Chapter 3: The Relationship Between Antibiotic and Metal Resistance Co-Occurrence and the Spread of Multidrug-Resistant Nontyphoidal Salmonella

INTRODUCTION

Antibiotic resistant *Salmonella enterica subspecies enterica* is classified as a serious threat to public health by the Centers for Disease Control and Prevention in the United States (1). *S. enterica* can rapidly disseminate antimicrobial resistance (AMR) genes horizontally, given the pathogen infects a wide range of hosts. Furthermore, *S. enterica* is common to both human and agricultural animals. It has been shown that sub-clinical doses of veterinary antibiotics in feed can promote the acquisition of resistance to clinically relevant antibiotics for human use (2). To combat the spread of AMR in bacteria, metals such as copper and zinc have been proposed and adopted by some producers as an alternative to antibiotics (3, 4). One supplement, copper sulphate, has been proposed as a growth promoter in swine feed since at least 1961 (5). The interest in metals as alternatives to antibiotics as growth promotors is obvious: metals are not antibiotics, thus curbing the spread of antibiotic resistance, copper is relatively inexpensive, and the growth promoting benefits, namely feed conversion efficiency, are retained when using the metal. When the European Union decided to ban the use of antibiotics as growth promoters in animal feed in 2006, many pig producers looked to copper sulphate as an alternative. However, it has been shown that pharmacological doses of copper sulphate in feed increases antibiotic resistance of *Escherichia coli* in pigs (6). A

similar trend has been observed in strains of *Enterococcus faecium resistant to copper sulphate which contain resistance to macrolides and glycopeptides on the same conjugative plasmid as the metal resistance* (7). Compounding the issue are reports demonstrating the cohabitation of metal and antibiotic resistance genes in linkage groups that are co-selected (plasmids and transposons) (8-10).

Co-occurrence of metal and antibiotic resistance genes in S. enterica have been documented with a focus on *S. enterica* I 4,[5],12:i:- (11-14). Given Salmonellas unique ability to act as a zoonotic disseminator of AMR and metal resistance genes, we sought to identify co-occurrence between metal and antibiotic resistance genes in a query set of >56,000 genomes. The query set contains 20 major serovars of S. enterica. We specifically focused upon metal resistance genes that are plasmid bound to focus our analysis on genes that could be horizontally transmitted.

MATERIALS AND METHODS

Salmonella enterica genome assembly acquisition

Genome assemblies and GenBank files for *S. enterica* isolates were downloaded from Genome on NCBI (ftp.ncbi.nih.gov::genomes/all/GCA/, October 10, 2019). Resultant GenBank files were parsed and only assemblies containing a collection date were included for further analysis. Downloaded assemblies were removed from the data set if the contig number was greater than 300. Genome assemblies were in silico serotyped using SISTR (15). This was done to ensure consistent serovar annotation and limit the impact of improper or mislabeled serotypes. The top 20 serovars in the dataset were used for co-occurrence identification. 20 serovars were chosen as each contained greater than 1,000 genomes, save for S. enterica Javiana, which contained n = 995.

Identification of Plasmid Replicons, Metal and Antibiotic Resistance Homologues

Plasmid replicons were identified in genome assemblies using the tool Abricate (16) against the PlasmidFinder (17) database. Hits were defined as sequences with >= 90% percent identity and >= 60% query coverage.

Open-reading frames (ORF) in assemblies were predicted using Prodigal (18). Metal resistance homologues were identified in the resultant ORF files using GhostZ (19) to compare the ORF against the experimentally verified BacMet database (version 2.0) (20). ORF were considered homologues when: e-value <= 1 E-6, percent identity >= 90 % and percent coverage >= 60%. As we aimed to identify metal resistance under possible recombination, only homologues marked “PLASMID” in the BacMet database were considered.

Antibiotic resistance homologues were identified in an analogous manner to the metal resistance genes, however, ORF were queried against the NCBI antimicrobial resistance database (BioProject: PRJNA313047) using the same parameters.

Co-Occurrence Identification

Plasmid replicons, metal and antibiotic resistance homologues from each assembly were compiled into a binary matrix. A co-occurrence matrix was generated by taking the cross product of the original binary matrix. Simply put, if two genes are contained in the same assembly, they yield a value of 1. If genes are contained across multiple assemblies together, the number will increase accordingly. To identify possible co-occurrence between genes, the co-occurrence matrix was subjected to k-means clustering using clusters. The optimal number of clusters was determined by constructing an “elbow plot” (the sum of square distances against k number of clusters). Matrix compilation and k-

means clustering were conducted using R (R -core team). Igraph (21) was used to visualize the co-occurrence network. Nodes are scaled to edge weight number (larger nodes contain more co-occurrences).

Phylogeny and *S. enterica* I,4,[5],12:i:- Analysis

To compile a phylogeny estimation of 56,348 *S. enterica* genomes, MASH (22) sketches were drawn from each sample (10,000 sketches, kmer size of 21) and a pairwise distance matrix was generated. A neighbor-joining tree was constructed from the distance matrix using QuickTree (23).

Reads of European and US *S. enterica* I,4,[5],12:i:- and US *S. enterica* typhimurium and were downloaded from SRA hosted at NCBI (https://www.ncbi.nlm.nih.gov/sra). Resultant reads were mapped reads against a *S. enterica* Typhimurium DT104 NCTC 13348 reference strain downloaded from https://www.sanger.ac.uk/resources/downloads/bacteria/salmonella.html. Mapping and variant calling was conducted using Snippy (24) and only core variants were considered. A maximum likelihood from the SNP alignment was generated using IQTree (25) with a GTR+G4 substitution model and 1000 bootstraps. All phylogenies were visualized using the R package ggtree (26).

Reads were assembled into contigs using the package Shovill (https://github.com/tseemann/shovill) with a minimum contig length of 200 base pairs. Prodigal was used to predict ORF and plamid replicons, antibiotic and metal resistance genes were predicted in the same manner as the genomes of the large test set.

RESULTS

Broad Screen for Metal and Antibiotic Co-Occurrence in *S. enterica*

The aim of our study was to identify co-occurrence between metal and antibiotic resistance genes in *S. enterica* with the potential for recombination (plasmid bound). To begin the screening, approximately 81,000 genome assemblies were downloaded from NCBI, quality checked, and in silico serotyped. The twenty most abundant serotypes, all containing greater than or nearly 1,000 assemblies (*S. enterica* Javiana, n = 995), were taken for further analysis (table 1). The final dataset totaled 56,348 *S. enterica* assemblies and corresponding metadata is provided in supplementary table 1. Metal resistance, antibiotic resistance, and plasmid replicons were predicted from each assembly. Binary matrices describing homologue hits for each query type are provided in supplemental tables 2, 3, and 4. To gain understanding into the temporal distribution of metal and antibiotic resistance, the mean number of antibiotic and metal resistance genes were plotted against year of collection as shown in figure 1. Both resistance types show serovar specific trends. Many of the serovars show no trend between mean number of antibiotic and metal resistance genes over time. However, three serovars, *S. enterica* Kentucky, *S. enterica* I 4,[5],12:i:-, and *S. enterica* Infantis, have increasing antibiotic and metal resistance genes over time. The trend is most evident in *S. enterica* I 4,[5],12:i:-, with the mean number of metal resistance genes increasing nearly 3 fold in 15 years. S. enterica Schwarzengrund and S. enterica Senftenberg showed constant metal resistance with no trend observed to antibiotic resistance. S. enterica Enteritidis, the most abundant serovar, showed almost no antibiotic nor metal resistance homologues.

Co-Occurrence of Metal and Antibiotic Resistance

Co-occurrence between plasmid replicons (included as the focus was to identify possible recombinant resistance), metal resistance genes, and antibiotic resistance genes was identified by applying k-means clustering to a co-occurrence matrix (see methods). Three groups of co-occurrent genes were identified (figure 2A). To better visualize the groups identified by k-means clustering, the co-occurrence network was plotted, and groups were colored respective to figure 2A (figure 2B). Group 1 was the smallest cluster set containing only 11 genes comprising the *pco* and *sil* operons. Said operons confer resistance to copper and silver respectively. Group 2 represents most plasmid replicons, metal resistance genes, and antibiotic resistance genes. Little co-occurrence is observed in group 2 save for two plasmid replicons (IncFIB(S)_1 and IncFII(S)_1). The small node sizes of other constituents in group 2 suggests that gene turnover may be high in the associated plasmids. The majority of antibiotic and metal resistance genes are not co-occurrent in the dataset of *S. enterica* genomes. However, co-occurrence was identified in group 3. Homologues in group 3 confer resistance to metals arsenic and mercury, antibiotic classes of beta-lactams, sulfonamides, tetracyclines, and contains one plasmid replicon: IncQ1_1. The top down approach and simple k-means clustering was able to identify that co-occurrence between metal and antibiotic resistance is occurring in some genomes of S. enterica.

However, it also showed that the effect is only occurring in a small subset of genes, and that most are not carried together. The full list of homologues in groups 1 and 3 are presented in table 2.

To better visualize the distribution of group 1 and 3 gene clusters (moderate to high co-occurrent genes), a presence-absence matrix of group 1 and 3 genes was plotted against the phylogeny of *S. enterica* I 4,[5],12:i:-, S. enterica Infantis, S. enterica Kentucky, S. enterica Schwarzengrund, and S. enterica Senftenberg (serovars that showed trends in figure 1) as shown in figure 3. Group 1 (sil and pco) genes are broadly distributed among S. enterica I 4,[5],12:i:-, S. enterica Kentucky, S. enterica Schwarzengrund, and S. enterica Senftenberg, but are largely absent from S. enterica Infantis. However, group 1 genes only confer resistance to the metals silver and copper, and not antibiotics. Only *S. enterica* I 4,[5],12:i:- assemblies consistently contained loci from group 3, the co-occurrent group of metal and antibiotic resistance genes. Therefore, we decided to focus the study upon *S. enterica* I 4,[5],12:i:- and the possible mechanisms of the co-occurrence. Recent research has proposed the addition of metals, namely silver and copper, to medical devices, food pack

S. enterica I,4,[5],12:i:- Metal and Antibiotic Co-Occurrence

S. enterica I 4,[5],12:i:- was the only serovar to contain group 3 genes in a substantive manner. The initial dataset was compiled from assembled genomes in order to facilitate speed. However, the differing assembly mechanisms and lack of SNP based phylogeny limits the accuracy. To remedy this, 1,455 sequencing reads of US S. enterica I 4,[5],12:i:-, 1,447 sequencing reads of US S. enterica Typhimurium and 329 sequencing reads of European S. enterica I 4,[5],12:i:- were compiled into a SNP phylogeny against a reference S. typhimurium DT104 assembly. The addition of the new data set allows for a validation step against the initial screening. Plasmid replicons, metal and antibiotic resistance were annotated in the second dataset exactly following the

protocol for the large query set. A presence-absence matrix describing group 1, 3, and major plasmid replicons is plotted against the phylogeny of S. enterica I 4,[5],12:i:- and S. enterica Typhimurium in figure 4. Confirming the initial screening, group 1 and 3 genes were carried in S. enterica I 4,[5],12:i:- but were mostly absent from S. Typhimurium. Two major clades of S. enterica I 4,[5],12:i:- are observed: one comprising solely of US isolates and a mixed clade of US and European genomes. Notably, the mixed clade of US and European isolates contain both group 1 and 3 as well as the IncQ1_1 plasmid replicon. The S. enterica I 4,[5],12:i:- clade comprised solely of US genomes did not contain group 1 and 3 genes. To this end, the co-occurrence of metal and antibiotic resistance is not widespread throughout S. enterica, but rather is localized to one clade of S. enterica I 4,[5],12:i:- . The metadata for the detailed analysis of S. enterica I 4,[5],12:i:- and S. enterica Typhimurium is provided in supplemental table 5.

DISCUSSION

In this work, we identified co-occurrence of metal and antibiotics in *S. enterica* I 4,[5],12:i:- by screening a broad set of 20 serovars totaling more than 56,000 genomes. The top-down approach allows for a non-biased and comprehensive approach compared to starting with metal resistant isolates and working upwards towards the serovar level. Ultimately, our hypothesis was incorrect as co-occurrence between metal and antibiotic resistance is not widespread in S. enterica. Rather, the phenomenon is limited to one clade of *S. enterica* I 4,[5],12:i:-. Metal and antibiotic resistance co-occurrence may occur in other serovars not included in this study; however we chose to limit our analysis to serovars with > 1,000 samples to ensure that outlier strains would not have a drastic

impact on the output. One unexpected outcome was the co-occurrence of pco and sil operons in multiple serovars.

The pco and sil operons are associated in S. enterica with Salmonella Genomic Island 4 (SGI-4). SGI-4 is a chromosomal island containing the *pco* and *sil* operons that may be excised from the chromosome by mitomycin C, and oxygen tension related stress (11). SGI-4 was first designated in 2016 (14)(see addendum) and is largely associated with *S. enterica* I 4,[5],12:i:- (11). Indeed it has been stated that SGI-4 has only been discovered in I 4,[5],12:i:- (27). Our results indicate the genomic island may be more widespread than believed and may be contained by: S. enterica I 4,[5],12:i:-, S. enterica Kentucky, S. enterica Schwarzengrund, and S. enterica Senftenberg. The resistance to copper and silver may increase the clinical relevance of the serovars. A burn ward identified S. enterica Senftenberg as a causative agent of a burn ward infection (28). The isolated strains were resistant to the silver sulfadiazine that was applied daily to the burn patients. It has been proposed that copper and silver can be used in medical implant design as antimicrobial agents (29, 30) which may create novel infectious niches for metal resistant strains.

The serovar specific nature of group 3 genes (metal and antibiotic co-occurrence) is consistent with earlier reports (12, 31, 32). Interestingly, the co-occurrence is not a characteristic of *S. enterica* I 4,[5],12:i:-., but rather one clade that appears to have European origins. Previous work has demonstrated that a clade *S. enterica* I 4,[5],12:i:- circulating the US is descendent from multi-drug resistant strains in Europe (33). The inclusion of the IncQ1_1 plasmid replicon reinforces the claim as the gene was the most identified plasmid replicon of *S. enterica* I 4,[5],12:i:- in Italian isolates (12). A plasmid

harboring resistance to metals and antibiotics has been documented in an Australian *S. enterica* I 4,[5],12:i:- isolated from pig feces (13). However, the plasmid was large (~ 275 kb) and annotated as a IncHI2 class plasmid, not an IncQ class as we report here. Additionally, it was found that the plasmid was able to conjugate to other species (13). IncQ plasmids are unable to self-transfer and rely upon "helper plasmids" for transmission between bacterial cells (34). The lack of other plasmid replicons in large numbers in the mixed US and European clade of *S. enterica* I 4,[5],12:i:-, and isolation of group 3 genes in the clade, suggests the plasmid and group 3 genes are being transmitted vertically rather than horizontally. Furthermore, IncQ class plasmids are evolutionary stable with one study documenting the presence of the plasmid in the environment over a 30 year period (35). Taken together, we propose that the method of metal and antibiotic co-occurrence in the US and European monophasic strains studied here differs from the large IncHI2 plasmid isolated from an Australian strain of *S. enterica* I 4,[5],12:i:-. Complicating the narrative is the difficulty in accurately constructing plasmid contigs from short-read sequencing (36). Isolating plasmids from a large number of *S. enterica* I 4,[5],12:i:- containing the IncQ1_1 and group 3 genes and sequencing with a long-read platform such as PacBio is needed to confirm our results.

REFERENCES

1. CDC. 2019 Antibiotic Resistance Threats in the United States.

2. Singh AK, Bhunia AK. 2019. Animal-Use Antibiotics Induce Cross-Resistance in Bacterial Pathogens to Human Therapeutic Antibiotics. Current Microbiology 76:1112-1117.

3. Liu Y, Espinosa CD, Abelilla JJ, Casas GA, Lagos LV, Lee SA, Kwon WB, Mathai JK, Navarro DMDL, Jaworski NW, Stein HH. 2018. Non-antibiotic feed additives in diets for pigs: A review. Animal Nutrition 4:113-125.

4. Carpenter CB, Woodworth JC, DeRouchey JM, Tokach MD, Goodband RD, Dritz SS, Wu F, Usry JL. 2018. Effects of increasing copper from tri-basic copper chloride or a copper-methionine chelate on growth performance of nursery pigs1,2. Translational Animal Science 3:369-376.

5. Lucas IAM, Livingstone RM, McDonald I. 1961. Copper sulphate as a growth stimulant for pigs: effect of level and purity. Animal Science 3:111-119.

6. Zhang Y, Zhou J, Dong Z, Li G, Wang J, Li Y, Wan D, Yang H, Yin Y. 2019. Effect of Dietary Copper on Intestinal Microbiota and Antimicrobial Resistance Profiles of Escherichia coli in Weaned Piglets. Frontiers in Microbiology 10:2808.

7. Hasman H, Aarestrup FM. 2002. tcrB, a Gene Conferring Transferable Copper Resistance in Enterococcus faecium: Occurrence, Transferability, and Linkage to Macrolide and Glycopeptide Resistance. Antimicrobial Agents and Chemotherapy 46:1410.

8. Baker-Austin C, Wright MS, Stepanauskas R, McArthur JV. 2006. Co-selection of antibiotic and metal resistance. Trends in Microbiology 14:176-182.

9. Seiler C, Berendonk T. 2012. Heavy metal driven co-selection of antibiotic resistance in soil and water bodies impacted by agriculture and aquaculture. Frontiers in Microbiology 3:399.

10. Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DGJ. 2015. Co-occurrence of resistance genes to antibiotics, biocides and metals reveals novel insights into their co-selection potential. BMC Genomics 16:964.

11. Branchu P, Charity OJ, Bawn M, Thilliez G, Dallman TJ, Petrovska L, Kingsley RA. 2019. SGI-4 in Monophasic Salmonella Typhimurium ST34 Is a Novel ICE That Enhances Resistance to Copper. Frontiers in Microbiology 10:1118.

12. Mastrorilli E, Pietrucci D, Barco L, Ammendola S, Petrin S, Longo A, Mantovani C, Battistoni A, Ricci A, Desideri A, Losasso C. 2018. A Comparative Genomic Analysis Provides Novel Insights Into the Ecological Success of the Monophasic Salmonella Serovar 4,[5],12:i. Frontiers in microbiology 9:715-715.

13. Dyall-Smith ML, Liu Y, Billman-Jacobe H. 2017. Genome Sequence of an Australian Monophasic Salmonella enterica subsp. enterica Typhimurium Isolate (TW-Stm6) Carrying a Large Plasmid with Multiple Antimicrobial Resistance Genes. Genome announcements 5:e00793-17.

14. Liljana P, Alison EM, Manal A, Priscilla B, Simon RH, Thomas C, Hopkins KL, Underwood A, Antonia AL, Andrew JP, Mary B, John W, Julian P, Gordon D, Robert D, Robert AK. 2016. Microevolution of Monophasic Salmonella Typhimurium during Epidemic, United Kingdom, 2005–2010. Emerging Infectious Disease journal 22:617.

15. Yoshida CE, Kruczkiewicz P, Laing CR, Lingohr EJ, Gannon VPJ, Nash JHE, Taboada EN. 2016. The Salmonella In Silico Typing Resource (SISTR): An Open Web-Accessible Tool for Rapidly Typing and Subtyping Draft Salmonella Genome Assemblies. PLOS ONE 11:e0147101.

16. Seemann T. 2018. ABRicate, v0.8.0. GitHub. https://github.com/tseemann/abricate.

17. Carattoli A, Zankari E, García-Fernández A, Voldby Larsen M, Lund O, Villa L, Møller Aarestrup F, Hasman H. 2014. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrobial agents and chemotherapy 58:3895-3903.

18. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC bioinformatics 11:119-119.

19. Suzuki S, Kakuta M, Ishida T, Akiyama Y. 2014. Faster sequence homology searches by clustering subsequences. Bioinformatics 31:1183-1190.

20. Pal C, Bengtsson-Palme J, Rensing C, Kristiansson E, Larsson DGJ. 2014. BacMet: antibacterial biocide and metal resistance genes database. Nucleic acids research 42:D737-D743.

21. Csardi G, Nepusz T. 2005. The Igraph Software Package for Complex Network Research. InterJournal Complex Systems:1695.

22. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM. 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology 17:132.

23. Howe K, Bateman A, Durbin R. 2002. QuickTree: building huge Neighbour-Joining trees of protein sequences. Bioinformatics 18:1546-1547.

24. Seemann T. 2015. snippy: fast bacterial variant calling from NGS reads, https://github.com/tseemann/snippy.

25. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution 32:268-274.

26. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 8:28-36.

27. Branchu P, Bawn M, Kingsley RA. 2018. Genome Variation and Molecular Epidemiology of Salmonella enterica Serovar Typhimurium Pathovariants. Infection and Immunity 86:e00079-18.

28. Nair D, Gupta N, Kabra S, Ahuja RB, Prakash SK. 1999. Salmonella senftenberg: a new pathogen in the burns ward. Burns 25:723-7.

29. Bergemann C, Zaatreh S, Wegner K, Arndt K, Podbielski A, Bader R, Prinz C, Lembke U, Nebe JB. 2017. Copper as an alternative antimicrobial coating for implants - An in vitro study. World journal of transplantation 7:193-202.

30. Besinis A, Hadi SD, Le HR, Tredwin C, Handy RD. 2017. Antibacterial activity and biofilm inhibition by surface modified titanium alloy medical implants following application of silver, titanium dioxide and hydroxyapatite nanocoatings. Nanotoxicology 11:327-338.

31. Mourão J, Marçal S, Ramos P, Campos J, Machado J, Peixe L, Novais C, Antunes P. 2016. Tolerance to multiple metal stressors in emerging non-typhoidal MDR Salmonella serotypes: a relevant role for copper in anaerobic conditions. Journal of Antimicrobial Chemotherapy 71:2147-2157.

32. Billman-Jacobe H, Liu Y, Haites R, Weaver T, Robinson L, Marenda M, Dyall-Smith M. 2018. pSTM6-275, a Conjugative IncHI2 Plasmid of Salmonella enterica That Confers Antibiotic and Heavy-Metal Resistance under Changing Physiological Conditions. Antimicrobial Agents and Chemotherapy 62:e02357-17.

33. Elnekave E, Hong S, Mather AE, Boxrud D, Taylor AJ, Lappi V, Johnson TJ, Vannucci F, Davies P, Hedberg C, Perez A, Alvarez J. 2017. Salmonella enterica Serotype 4,[5],12:i:- in Swine in the United States Midwest: An Emerging Multidrug-Resistant Clade. Clinical Infectious Diseases 66:877-885.

34. Rawlings DE, Tietze E. 2001. Comparative biology of IncQ and IncQ-like plasmids. Microbiology and molecular biology reviews : MMBR 65:481-496.

35. Yau S, Liu X, Djordjevic SP, Hall RM. 2010. RSF1010-Like Plasmids in Australian Salmonella enterica Serovar Typhimurium and Origin of Their sul2-strA-strB Antibiotic Resistance Gene Cluster. Microbial Drug Resistance 16:249-252.

36. Arredondo-Alonso S, Willems RJ, van Schaik W, Schurch AC. 2017. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. Microb Genom 3:e000128.

Figures and Tables

Table 1. List of serovars and number of genome assemblies, median assembly length, and date range of assemblies used to identify co-occurrence of metal and antibiotic resistance.

Serotype	n	Median Length (Mb)	Date Range
Enteritidis	11325	4.69	1950-2019
Typhimurium	8561	4.91	1958-2019
Kentucky	4131	4.92	1972-2019
Infantis	3866	4.94	1971-2019
I 4,[5],12:i:-	3657	4.95	1985-2019
Newport	3646	4.76	1975-2019
Typhi	2736	4.74	1958-2019
Heidelberg	2454	4.89	1979-2019
Montevideo	1841	4.65	1997-2019
Agona	1718	4.82	1952-2019
Muenchen	1641	4.79	1987-2019
Saintpaul	1570	4.79	1974-2019
Anatum	1560	4.73	1993-2019
Senftenberg	1329	4.81	2001-2019
Schwarzengrund	1130	4.81	2000-2019
Mbandaka	1083	4.75	2000-2019
Braenderup	1044	4.69	1999-2019
Hadar	1038	4.74	1988-2019
Derby	1023	4.87	1986-2019
Javiana	995	4.61	1995-2019

Table 2. Major groups of co-occurrent genes identified by k-means clustering.

Cluster	Gene	Class	Subclass
1	pcoA	Copper	Copper
	pcoB	Copper	Copper
	pcoC	Copper	Copper
	pcoD	Copper	Copper
	pcoR	Copper	Copper
	pcoS	Copper	Copper
	silA	Silver	Silver
	silB	Silver	Silver
	silC	Silver	Silver
	silE	Silver	Silver
	silF	Silver	Silver
3	aph(3")-Ib	Aminoglycoside	Streptomycin
	aph(6)-Id	Aminoglycoside	Streptomycin
	arsA	Arsenic	Arsenite
	arsB	Arsenic	Arsenite
	arsC	Arsenic	Arsenate
	arsD	Arsenic	Arsenite
	arsR	Arsenic	Arsenite
	blaTEM	Beta-Lactam	Beta-Lactam
	IncQ1_1	Plasmid Replicon	NA
	merR	Mercury	Mercury
	merT	Mercury	Mercury
	qacE	Quaternary Ammonium	Quaternary Ammonium
	qacEdelta1	Quaternary Ammonium	Quaternary Ammonium
	sul1	Sulfonamide	Sulfonamide
	sul2	Sulfonamide	Sulfonamide
	tet(A)	Tetracycline	Tetracycline
	tet(B)	Tetracycline	Tetracycline

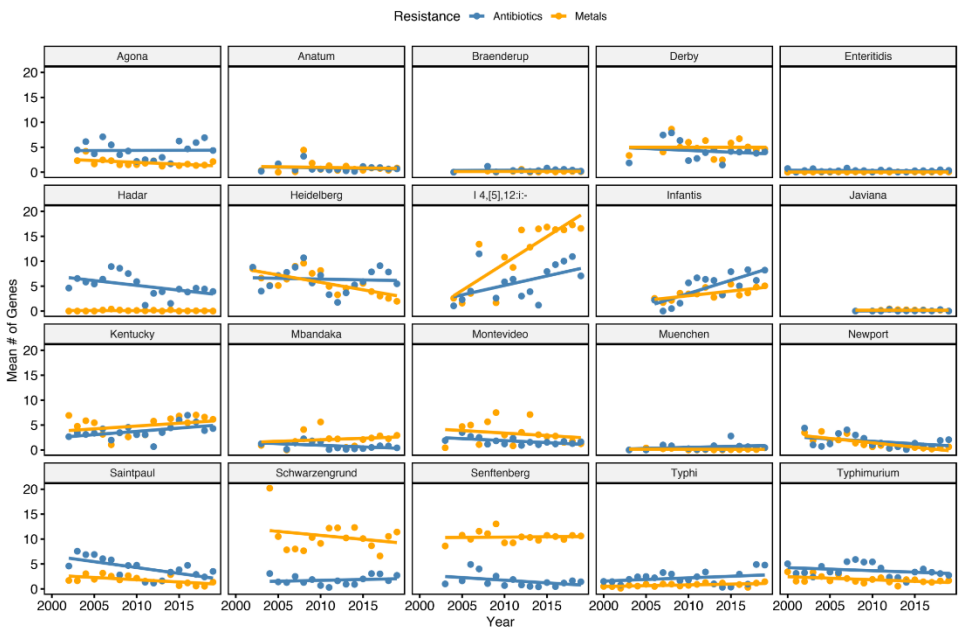


Figure 1. Temporal distribution of antibiotic and metal resistance genes by S. enterica serovar. The mean number of antibiotic and resistance homologues identified in each serovar is plotted against the date of collection. Only dates that contained n > 10 genomes were considered to control for outliers.

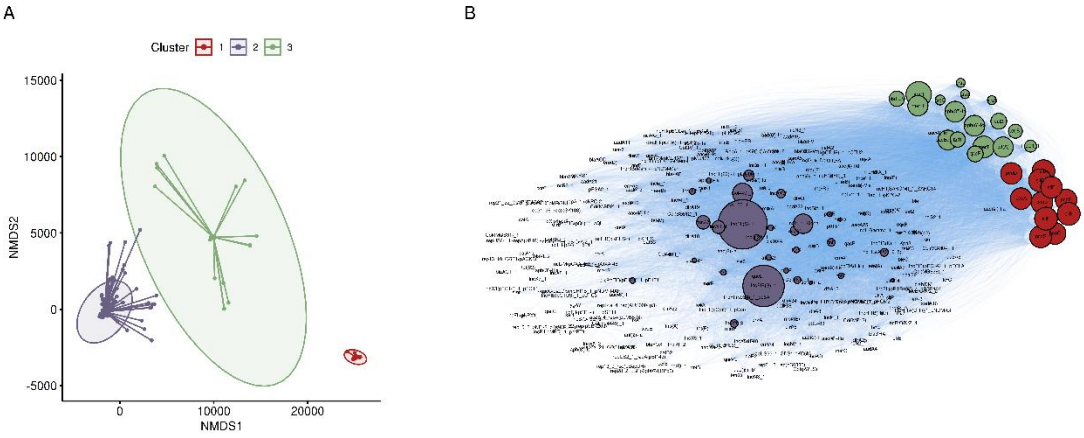


Figure 2. Co-occurrent gene clusters identified in S. enterica genomes. (A) Non-metric Multi-dimensional Scaling plot of a co-occurrence matrix generated from plasmid replicons, antibiotic and metal resistance homologues. K-means clustering was used to group genes. An elbow plot was used to pick the optimal number of clusters. Clusters are drawn as a star plot with lines connecting each gene to the center of the cluster. Gaussian distributions were drawn around each cluster. (B) Network representation of co-occurrence network. Nodes are scaled to edge weight (number of connections) and are colored consistent to A.

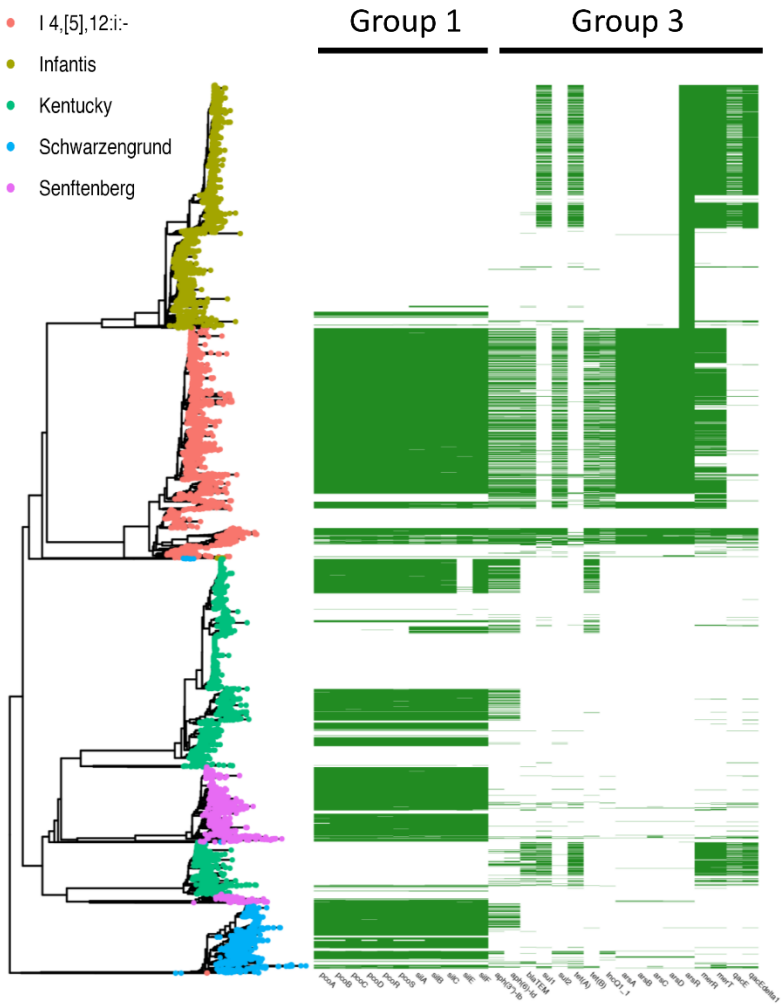


Figure 3. Distribution of group 1 and 3 genes among *S. enterica* I 4,[5],12:i:-, S. enterica Infantis, S. enterica Kentucky, S. enterica Schwarzengrund, and S. enterica Senftenberg. A prescence-absence matrix of group 1 and 3 genes is plotted against a neighbor-joining tree constructed from pairwise distance matrix. Group 1 (pco and sil operons) is found in S. enterica Kentucky, S. enterica Schwarzengrund, and S. enterica Senftenberg but is mostly absent from S. enterica Infantis. Group 3, which contains metal and antibiotic resistance genes, is almost exclusive to *S. enterica* I 4,[5],12:i:-.

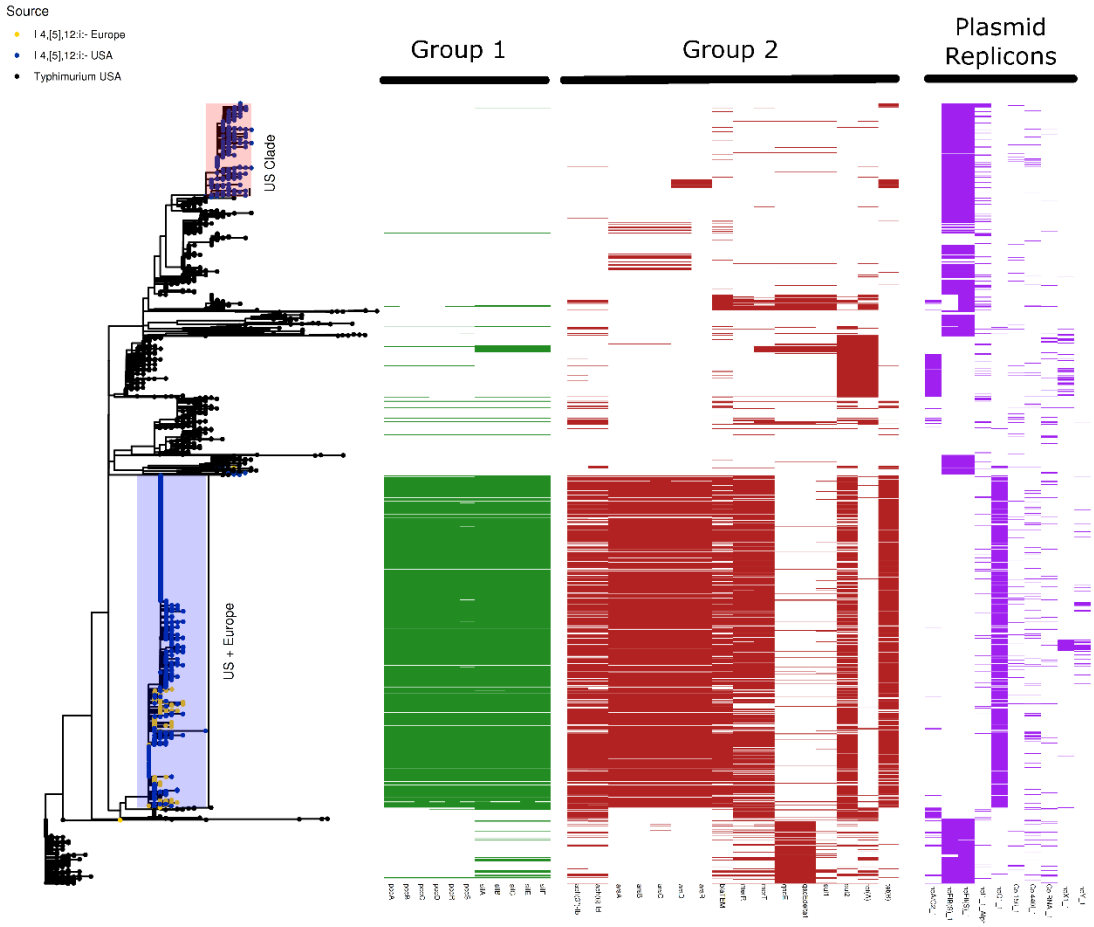


Figure 4. Distribution of group 1, group3, and major plasmid replicons in *S. enterica* I 4,[5],12:i:-. A presence-absence matrix for each gene group is plotted against a maximum-likelihood tree of S. enterica Typhimurium and *S. enterica* I 4,[5],12:i:- rooted a reference S. enterica Typhimurium DT104 assembly. Two major clades of *S. enterica* I 4,[5],12:i:- are observed, however only the clade containing both US and European isolates contains both group 1 and group 3 genes as well as the IncQ1_1 plasmid replicon.